Slice inverse regression with score functions

Dmitry Babichev, Francis Bach

Electronic Journal of Statistics
ISSN: 1935-7524

Slice inverse regression with score functions

Dmitry Babichev and Francis Bach

INRIA - Sierra project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France,
e-mail: dmitry.babichev@inria.fr; francis.bach@inria.fr

Abstract: We consider non-linear regression problems where we assume that the response depends non-linearly on a linear projection of the covariates. We propose score function extensions to sliced inverse regression problems, both for the first- order and second-order score functions. We show that they provably improve estimation in the population case over the non-sliced versions and we study finite sample estimators and their consistency given the exact score functions. We also propose to learn the score function as well, in two steps, i.e., first learning the score function and then learning the effective dimension reduction space, or directly, by solving a convex optimization problem regularized by the nuclear norm. We illustrate our results on a series of experiments.

MSC 2010 subject classifications: Primary 62J02, 62G20; secondary 62G05.
Keywords and phrases: projection pursuit, dimension reduction, non-linear regression, slice inverse regression.

Contents

1. Introduction

Non-linear regression and related problems such as non-linear classification are core important tasks in machine learning and statistics. In this paper, we consider a random vector $x \in \mathbb{R}^d$, a random response $y \in \mathbb{R}$, and a regression model of the form

$$y = f(x) + \varepsilon, \tag{1.1}$$

which we want to estimate from n independent and identically distributed (i.i.d.) observations (x_i, y_i), $i = 1, \dots, n$. Our goal is to estimate the function f from these data. A traditional key difficulty in this general regression problem is the lack of parametric assumptions regarding the functional form of f, leading to a problem of *non-parametric* regression. This is often tackled by searching implicitly or explicitly a function f within an infinite-dimensional vector space.

While several techniques exist to estimate such a function, e.g., kernel methods, local-averaging, or neural networks (see, e.g., [14]), they also suffer from the *curse of dimensionality*, that is, the rate of convergence of the estimated function to the true function (with any relevant performance measure) can only decrease as a small power of n, and this power cannot be larger than a constant divided by d. In other words, the number n of observations for any level of precision is exponential in dimension.

A classical way of by-passing the curse of dimensionality is to make extra assumptions regarding the function to estimate, such as the dependence on a lower unknown low-dimensional subspace, such as done by projection pursuit or neural networks. More precisely, throughout the paper, we make the following assumption:

(A1) For all $x \in \mathbb{R}^d$, we have $f(x) = g(w^\top x)$ for a certain matrix $w \in \mathbb{R}^{d \times k}$ and a function $g : \mathbb{R}^k \to \mathbb{R}$. Moreover, $y = f(x) + \varepsilon$ with ε independent of x with zero mean and finite variance.

The subspace of $\mathbb{R}^d$ spanned by the k columns $w_1, \dots, w_k \in \mathbb{R}^d$ of w has dimension less than or equal to k, and is often called the *effective dimension reduction* (e.d.r.) space. The model above is often referred to as a *multiple-index model* [38]. We will always make the assumption that the e.d.r. space has exactly rank k, that is the matrix w has rank k (which implies that $k \leqslant d$).

Given w, estimating g may be done by any technique in non-parametric regression, with a convergence rate which requires a number of observations n to be exponential in k, with methods based on local averaging (e.g., Nadaraya-Watson estimators) or on least-squares regression [see, e.g., 14, 29]. Given the non-linear function g, estimating w is computationally difficult because the resulting optimization problem may not be convex and thus leads to several local

minima. The difficulty is often even stronger since one often wants to estimate *both* the function g and the matrix w.

Our main goal in this paper is to estimate the matrix w, with the hope of obtaining a convergence rate where the inverse power of n will now be proportional to k and not d. Note that the matrix w is only identifiable up to a (right) linear transform, since only the subspace spanned by its column is characteristic.

Method of moments vs. optimization. This multiple-index problem and the goal of estimating w only can be tackled from two points of views: (a) the method of moments, where certain moments are built so that the effect of the unknown function g disappears [6, 23], a method that we follow here and describe in more details below. These methods rely heavily on the model being correct, and in the instances that we consider here lead to provably polynomial-time algorithms (and most often linear in the number of observations since only moments are computed). In contrast, (b) optimization-based methods use implicitly or explicitly non-parametric estimation, e.g., using local averaging methods to design an objective function that can be minimized to obtain an estimate of w [35, 13]. The objective function is usually non-convex and gradient descent techniques are used to obtain a local minimum. While these procedures offer no theoretical guarantees due to the potential unknown difficulty of the optimization problem, they often work well in practice, and we have observed this in our experiments.

In this paper, we consider and improve a specific instantiation of the method of moments, which partially circumvents the difficulty of joint estimation by estimating w directly without the knowledge of g. The starting point for this method is the work by Brillinger [6], which shows, as a simple consequence of Stein's lemma [26], that if the distribution of x is Gaussian, **(A1)** is satisfied with $k = 1$ (e.d.r. of dimension one, e.g., a single-index model), and the input data have zero mean and identity covariance matrix, then the expectation $\mathbb{E}(yx)$ is proportional to w. Thus, a certain expectation, which can be easily approximated given i.i.d. observations, *simultaneously* eliminates g and reveals w.

While the result above provides a very simple algorithm to recover w, it has several strong limitations: (a) it only applies to normally distributed data x, or more generally to elliptically symmetric distributions [7], (b) it only applies to $k = 1$, and (c) in many situations with symmetries, the proportionality constant is equal to zero and thus we cannot recover the vector w. This has led to several extensions in the statistical literature which we now present.

Using score functions. The use of Stein's lemma with a Gaussian random variable can be directly extended using the score function $\mathcal{S}_1(x)$ defined as the negative gradient of the log-density, that is, $\mathcal{S}_1(x) = -\nabla \log p(x) = \frac{-1}{p(x)} \nabla p(x)$, which leads to the following assumption:

(A2) The distribution of x has a strictly positive density $p(x)$ which is differentiable with respect to the Lebesgue measure, and such that $p(x) \to 0$ when $\|x\| \to +\infty$.

We will need the score to be sub-Gaussian to obtain consistency results. Given Assumption **(A2)**, then [28] showed, as a simple consequence of integration by parts, that, for $k = 1$ and if Assumption **(A1)** is satisfied, then $\mathbb{E}(y\mathcal{S}_1(x))$ is proportional to w, for all differentiable functions g, with a proportionality constant that depends on w and ∇g. This leads to the "average derivative method" (ADE) and thus replaces the Gaussian assumption by the existence of a differentiable log-density, which is much weaker. This however does not remove the restriction $k = 1$, which can be done in two ways which we now present.

Sliced inverse regression. Given a normalized Gaussian distribution for x (or any elliptically symmetric distribution), then, if **(A1)** is satisfied, almost surely in y, the conditional expectation $\mathbb{E}(x|y)$ happens to belong to the e.d.r. subspace. Given several distinct values of y, the vectors $\mathbb{E}(x|y)$ or any estimate thereof, will hopefully span the entire e.d.r. space and we can recover the entire matrix w, leading to "slice inverse regression" (SIR), originally proposed by Li and Duan [23], Duan and Li [12] and Li [21]. This allows the estimation with $k > 1$, but this is still restricted to Gaussian data. In this paper, we propose to extend SIR by the use of score functions to go beyond elliptically symmetric distributions, and we show that the new method combining SIR and score functions is formally better than the plain ADE method.

From first-order to second-order moments. Another line of extension of the simple method of Brillinger [6] is to consider higher-order moments, namely the matrix $\mathbb{E}(yxx^\top) \in \mathbb{R}^{d\times d}$, which, with normally distributed input data x and, if **(A1)** is satisfied, will be proportional (in a particular form to be described in Section 2.2) to the Hessian of the function g, leading to the method of "principal Hessian directions" (PHD) from Li [22]. Again, $k > 1$ is allowed (more than a single projection), but thus is limited to elliptically symmetric data. Janzamin, Sedghi and Anandkumar [19] proposed to used second-order score functions to go beyond this assumption. In order to define this new method, we consider the following assumption:

(A3) The distribution of x has a strictly positive density $p(x)$ which is twice differentiable with respect to the Lebesgue measure, and such that $p(x)$ and $\|\nabla p(x)\| \to 0$ when $\|x\| \to +\infty$.

Given **(A1)** and **(A3)**, then one can show [19] that $\mathbb{E}(y\mathcal{S}_2(x))$ will be proportional to the Hessian of the function g, where $\mathcal{S}_2(x) = \nabla^2 \log p(x) + \mathcal{S}_1(x)\mathcal{S}_1(x)^\top = \frac{1}{p(x)}\nabla^2 p(x)$, thus extending the Gaussian situation above where $\mathcal{S}_1$ was a linear function and $\mathcal{S}_2(x)$, up to linear terms, proportional to $xx^\top$.

In this paper, we propose to extend the method above to allow an SIR estimator for the second-order score functions, where we condition on y, and we show that the new method is formally better than the plain method of [19].

Learning score functions through score matching. Relying on score functions immediately raises the following question: is estimating the score

function (when not available) really simpler than our original problem of non-parametric regression? Fortunately, a recent line of work [1] has considered this exact problem, and formulated the task of density estimation directly on score functions, which is particularly useful in our context. We may then use the data, first to learn the score, and then to use the novel score-based moments to estimate w. We will also consider a direct approach that jointly estimates the score function and the e.d.r. subspace, by regularizing by a sparsity-inducing norm.

Fighting the curse of dimensionality. Learning the score function is still a non-parametric problem, with the associated curse of dimensionality. If we first learn the score function (through score matching) and then learn the matrix w, we will not escape that curse, while our direct approach is empirically more robust.

Note that Hristache, Juditsky and Spokoiny [16] suggested iterative improvements of the ADE method, using elliptic windows which shrink in the directions of the columns of w, stretch in all others directions and tend to flat layers orthogonal to w. Dalalyan, Juditsky and Spokoiny [11] generalize the algorithm to multi-index models and proved $\sqrt{n}$-consistency of the proposed procedure in the case when the structural dimension is not larger than 4 and weaker dependence for $d > 4$. In particular, they provably avoid the curse of dimensionality. Such extensions are outside the scope of this paper.

Contributions. In this paper, we make the following contributions:

- We propose score function extensions to sliced inverse regression problems, both for the first-order and second-order score functions. We consider the infinite sample case in Section 2 and the finite sample case in Section 3. They provably improve estimation in the population case over the non-sliced versions, while we study in Section 3 finite sample estimators and their consistency given the exact score functions.
- We propose in Section 4 to learn the score function as well, in two steps, i.e., first learning the score function and then learning the e.d.r. space parameterized by w, or directly, by solving a convex optimization problem regularized by the nuclear norm.
- We illustrate our results in Section 5 on a series of experiments.

2. Estimation with infinite sample size

In this section, we focus on the population situation, where we can compute expectations and conditional expectations exactly, while we focus on finite sample estimators with known score functions in Section 3 with consistency results in Section 3.3, and with learned score functions in Section 4.

2.1. SADE: Sliced average derivative estimation

Before presenting our new moments which will lead to the novel SADE method, we consider the non-sliced method, which is based on Assumptions **(A1)** and

(A2) and score functions (the method based on the Gaussian assumption will be derived later as corollaries). The ADE method is based on the following lemma:

Lemma 1 (ADE moment [28])**.** *Assume **(A1)**, **(A2)**, the differentiability of g and the existence of expectation $\mathbb{E}(g'(w^\top x))$. Then $\mathbb{E}(\mathcal{S}_1(x)y)$ is in the e.d.r. subspace.*

Proof. Since $y = f(x) + \varepsilon$, and ε is independent of x with zero mean, we have

$$\begin{aligned}\mathbb{E}(\mathcal{S}_1(x)y) &= \mathbb{E}(\mathcal{S}_1(x)f(x)) = \int_{\mathbb{R}^d} \frac{-\nabla p(x)}{p(x)} f(x)p(x)dx \\ &= -\int_{\mathbb{R}^d} \nabla p(x) f(x) dx = \int_{\mathbb{R}^d} p(x)\nabla f(x) dx \text{ by integration by parts,} \\ &= w \cdot \mathbb{E}(g'(w^\top x)),\end{aligned}$$

which leads to the desired result. Note that in the integration by parts above, the decay of $p(x)$ to zero for $\|x\| \to +\infty$ is needed. □

The ADE moment above only provides a single vector in the e.d.r. subspace, which can only potentially lead to recovery for $k = 1$, and only if $\mathbb{E}(g'(w^\top x)) \neq 0$, which may not be satisfied, e.g., if x has a a symmetric distribution and g is even.

We can now present our first new lemma, the proof of which relies on similar arguments as for SIR [23] but extended to score functions. Note that we do not require the differentiability of the function g.

Lemma 2 (SADE moment)**.** *Assume **(A1)** and **(A2)**. Then, $\mathbb{E}(\mathcal{S}_1(x)|y)$ is in the e.d.r. subspace almost surely (in y).*

Proof. We consider any vector $b \in \mathbb{R}^d$ in the orthogonal complement of the subspace $\mathrm{Span}\{w_1, \dots, w_k\}$. We need to show, that $b^\top \mathbb{E}(\mathcal{S}_1(x)|y) = 0$ with probability 1. We have by the law of total expectation

$$b^\top \mathbb{E}\big(\mathcal{S}_1(x)|y\big) = \mathbb{E}\Big(\mathbb{E}\big(b^\top \mathcal{S}_1(x)|w_1^\top x, \dots, w_k^\top x, y\big)|y\Big).$$

Because of Assumption **(A1)**, we have $y = g(w^\top x) + \varepsilon$ with ε independent of x, and thus

$$\mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x, y\big) = \mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x, \varepsilon\big) = \mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x\big) \text{ almost surely.}$$

This leads to

$$b^\top \mathbb{E}\big(\mathcal{S}_1(x)|y\big) = \mathbb{E}\big[\mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x\big)\big|y\big].$$

We now prove that almost surely $\mathbb{E}\big(b^\top \mathcal{S}_1(x)|w^\top x\big) = 0$, which will be sufficient to prove Lemma 2. We consider the linear transformation of coordinates: $\tilde{x} = \tilde{w}^\top x \in \mathbb{R}^d$, where $\tilde{w} = (w_1, \dots, w_k, w_{k+1}, \dots, w_d)$ is a square matrix with full rank obtained by adding a basis of the subspace orthogonal to the span of the k columns of w. Then, if $\tilde{p}$ is the density of $\tilde{x}$, we have $p(x) = (\det \tilde{w}) \cdot \tilde{p}(\tilde{x})$ and

thus $\nabla p(x) = (\det \tilde{w}) \cdot \tilde{w} \cdot \nabla \tilde{p}(\tilde{x})$ and $\tilde{b} = \tilde{w}^\top b = (0, \dots, 0, \tilde{b}_{k+1}, \dots, \tilde{b}_d) \in \mathbb{R}^d$ (because $b \perp \mathrm{Span}\{w_1, \dots, w_k\}$). The desired conditional expectation equals

$$\mathbb{E}\big(\tilde{b}^\top \tilde{\mathcal{S}}_1(\tilde{x}) | \tilde{x}_1, \dots, \tilde{x}_k\big),$$

since $\tilde{w}\tilde{\mathcal{S}}_1(\tilde{x}) = \dfrac{\tilde{w}\nabla\tilde{p}(\tilde{x})}{\tilde{p}(\tilde{x})} = \dfrac{\nabla p(x)}{p(x)} = \mathcal{S}_1(x)$ and hence $b^\top \mathcal{S}_1(x) = b^\top \tilde{w}\, \tilde{\mathcal{S}}_1(\tilde{x}) = \tilde{b}^\top \tilde{\mathcal{S}}_1(\tilde{x})$.

It is thus sufficient to show that $\int\limits_{\mathbb{R}^{d-k}} \tilde{b}^\top \tilde{\mathcal{S}}_1(\tilde{x}) \tilde{p}(\tilde{x}_1, \dots, \tilde{x}_d) d\tilde{x}_{k+1} \dots d\tilde{x}_d = 0$, for all $\tilde{x}_1, \dots, \tilde{x}_k$. We have

$$\int\limits_{\mathbb{R}^{d-k}} \tilde{b}^\top \tilde{\mathcal{S}}_1(\tilde{x}) \tilde{p}(\tilde{x}_1, \dots, \tilde{x}_d) d\tilde{x}_{k+1} \dots d\tilde{x}_d = \tilde{b}^\top \int\limits_{\mathbb{R}^{d-k}} \nabla \tilde{p}(\tilde{x}) \cdot d\tilde{x}_{k+1} \dots d\tilde{x}_d$$

$$= \sum_{j=k+1}^{d} \tilde{b}_j \cdot \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \cdots \int\limits_{-\infty}^{\infty} \frac{\partial \tilde{p}(\tilde{x})}{\partial \tilde{x}_j} d\tilde{x}_j \cdot \prod_{\substack{k+1 \leqslant t \leqslant d \\ t \neq j}} d\tilde{x}_t = 0,$$

because for any $j \in \{k+1, \dots, n\}$, $\int\limits_{-\infty}^{\infty} \dfrac{\partial \tilde{p}(\tilde{x})}{\partial \tilde{x}_j} d\tilde{x}_j = 0$ by Assumption **(A2)**. This leads to the desired result. □

The key differences are now that:

- Unlike ADE, by conditioning on different values of y, we have access to *several* vectors $\mathbb{E}(\mathcal{S}_1(x)|y) \in \mathbb{R}^d$.
- Unlike SIR, SADE does not require the linearity condition from [21] anymore and can be used with a smooth enough probability density.

In the population case, we will consider the following matrix (using the fact that $\mathbb{E}(\mathcal{S}_1(x)) = 0$):

$$\mathcal{V}_{1,\mathrm{cov}} = \mathbb{E}\big[\mathbb{E}(\mathcal{S}_1(x)|y)\mathbb{E}(\mathcal{S}_1(x)|y)^\top\big] = \mathrm{Cov}\big[\mathbb{E}(\mathcal{S}_1(x)|y)\big] \in \mathbb{R}^{d\times d},$$

which we will also denote $\mathbb{E}\big[\mathbb{E}(\mathcal{S}_1(x)|y)^{\otimes 2}\big]$, where for any matrix a, $a^{\otimes 2}$ denotes $aa^\top$. The matrix above is positive semi-definite, and its column space is included in the e.d.r. space. If it has rank k, then we can exactly recover the entire subspace by an eigenvalue decomposition. When $k = 1$, which is the only case where ADE may be used, the following proposition shows that if ADE allows to recover w, so is SADE.

We will also consider the other matrix (note the presence of the extra term y^2)

$$\mathcal{V}'_{1,\mathrm{cov}} = \mathbb{E}\big[y^2\mathbb{E}(\mathcal{S}_1(x)|y)\mathbb{E}(\mathcal{S}_1(x)|y)^\top\big],$$

because of its direct link with the non-sliced version. Note that we made the weak assumption of existence of matrices $\mathcal{V}_{1,\mathrm{cov}}$ and $\mathcal{V}'_{1,\mathrm{cov}}$, which is satisfied for majority of problems.

Proposition 1. *Assume* ***(A1)*** *and* ***(A2)****, with* $k = 1$*, as well as differentiability of* g *and existence of the expectation* $\mathbb{E}g'(w^\top x)$*. The vector* w *may be recovered from the ADE moment (up to scale) if and only if* $\mathbb{E}g'(w^\top x) \neq 0$*. If this condition is satisfied, then SADE also recovers* w *up to scale (i.e.,* $\mathcal{V}_{1,cov}$ *and* $\mathcal{V}'_{1,cov}$ *are different from zero).*

Proof. The first statement is a consequence of the proof of Lemma 1. If SADE fails, that is, for almost all y, $\mathbb{E}(\mathcal{S}_1(x)|y) = 0$, then $\mathbb{E}(\mathcal{S}_1(x)y|y) = 0$ which implies that $\mathbb{E}(\mathcal{S}_1(x)y) = 0$ and thus ADE fails. Moreover, we have, using operator convexity [33]:

$$\mathcal{V}'_{1,\mathrm{cov}} = \mathbb{E}\big[\mathbb{E}(y\mathcal{S}_1(x)|y)\mathbb{E}(y\mathcal{S}_1(x)|y)^\top\big] \succcurlyeq \big[\mathbb{E}(\mathbb{E}(y\mathcal{S}_1(x)|y))\big]\big[\mathbb{E}(\mathbb{E}(y\mathcal{S}_1(x)|y))\big]^\top =$$

$$= \big[\mathbb{E}(y\mathcal{S}_1(x))\big]\big[\mathbb{E}(y\mathcal{S}_1(x))\big]^\top,$$

showing that the new moment is dominating the ADE moment, which provides an alternative proof of the rank of $\mathcal{V}'_{1,\mathrm{cov}}$ being larger than one if $\mathbb{E}(y\mathcal{S}_1(x)) \neq 0$. □

Elliptically symmetric distributions. If x is normally distributed with mean vector μ and covariance matrix Σ, then we have $\mathcal{S}_1(x) = \Sigma^{-1}(x-\mu)$ and we recover the result from [23]. Note that the lemma then extends to all elliptical distributions of the form $\varphi(\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu))$, for a certain function $\varphi : \mathbb{R}^+ \to \mathbb{R}$. See [23].

Failure modes. In some cases, slice inverse regression does not span the entire e.d.r. space, because the inverse regression curve $\mathbb{E}(\mathcal{S}_1(x)|y)$ is degenerated. For example, this can occur, if $k = 1$, $y = h(w_1^\top x) + \varepsilon$, h is an even function and $w_1^\top x$ has a symmetric distribution around 0. Then $\mathbb{E}(\mathcal{S}_1(x)|y) \equiv 0$, and thus it is a poor estimation of the desired e.d.r. directions [10].

The second drawback of SIR occurs when we have a classification task, for example, $y \in \{0, 1\}$. In this case, we have only two slices (i.e., possible values of y) and SIR can recover only one direction in the e.d.r. space [9].

Li [22] suggested another way to estimate the e.d.r. space which can handle such symmetric cases: principal Hessian directions (PHD). However, this method uses the normality of the vector X. As in the SIR case, we can extend this method, using score functions to use it for any distribution, which we will refer to as SPHD, which we now present.

2.2. SPHD: Sliced principal Hessian directions

Before presenting our new moment which will lead to the SPHD method, we consider the non-sliced method, which is based on Assumptions **(A1)** and **(A3)** and score functions (the method based on the Gaussian assumption will be derived later as corollaries). The method of [19], which we refer to as "PHD+" is based on the following lemma (we reproduce the proof for readability):

Lemma 3 (second-order score moment (PHD+) [19]). *Assume **(A1)**,**(A3)**,twice differentiability of function g and existence of the expectation $\mathbb{E}(\nabla^2 g(w^\top x))$. Then $\mathbb{E}(\mathcal{S}_2(x)y)$ has a column space included in the e.d.r. subspace.*

Proof. Since $y = f(x) + \varepsilon$, and ε is independent of x, we have

$$\mathbb{E}\big(y\mathcal{S}_2(x)\big) = \mathbb{E}\big(f(x)\mathcal{S}_2(x)\big) = \int \frac{\nabla^2 p(x)}{p(x)} p(x) f(x) dx =$$

$$= \int \nabla^2 p(x) \cdot f(x) dx = \int p(x) \cdot \nabla^2 f(x) dx = \mathbb{E}\big[\nabla^2 f(x)\big],$$

using integration by parts and the decay of $p(x)$ and $\nabla p(x)$ for $\|x\| \to \infty$. This leads to the desired result since $\nabla^2 f(x) = w \nabla^2 g(w^\top x) w^\top$. This was proved by Li [22] for normal distributions. □

Failure modes. The method does not work properly if $\operatorname{rank}(\mathbb{E}(\nabla^2 g(w^\top x))) < k$. For example, if g is linear function, $\mathbb{E}(\nabla^2 g(w^\top x)) \equiv 0$ and the estimated e.d.r. space is degenerated. Moreover, the method fails in symmetric cases, for example, if g is an odd function with respect to any variable and $p(x)$ is even function, then $\operatorname{rank}(\mathbb{E}(\nabla^2 g(w^\top x))) < k$.

We can now present our second new lemma, the proof of which relies on similar arguments as for PHD [22] but extended to score functions (again no differentiability is assumed on g):

Lemma 4 (SPHD moment). *Assume **(A1)** and **(A3)**. Then, $\mathbb{E}(\mathcal{S}_2(x)|y)$ has a column space within the e.d.r. subspace almost surely.*

Proof. We consider any $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^d$ orthogonal to the e.d.r. subspace, and prove, that $a^\top \mathbb{E}\big(\mathcal{S}_2(x)|y\big) b = 0$. We use the same transform of coordinates as in the proof of Lemma 2: $\tilde{x} = \tilde{w}^\top x \in \mathbb{R}^d$. Then $\nabla^2 p(x) = \det(\tilde{w}) \cdot \tilde{w} \nabla^2 \tilde{p}(\tilde{x}) \tilde{w}^\top$, $\tilde{b} = \tilde{w}^\top b = (0, \dots, 0, \tilde{b}_{k+1}, \dots, \tilde{b}_d)$ and we will prove, that $\mathbb{E}\big(a^\top \mathcal{S}_2(x) b | w^\top x\big) = 0$ almost surely, and $\tilde{w}^\top \tilde{\mathcal{S}}_2(\tilde{x}) \tilde{w} = \dfrac{\tilde{w}^\top \nabla^2 \tilde{p}(\tilde{x}) \tilde{w}}{\tilde{p}(\tilde{x})} = \dfrac{\nabla^2 p(x)}{p(x)} = \mathcal{S}_2(x)$. It is sufficient to show, that for all $\tilde{x}_1, \dots, \tilde{x}_k$:

$$\int\limits_{\mathbb{R}^{d-k}} \tilde{a}^\top \cdot \tilde{\mathcal{S}}_2(\tilde{x}) \cdot \tilde{b} \cdot \tilde{p}(\tilde{x}_1, \dots, \tilde{x}_d) d\tilde{x}_{k+1} \dots d\tilde{x}_d = 0.$$

We have:

$$\int\limits_{\mathbb{R}^{d-k}} \tilde{a}^\top \cdot \tilde{\mathcal{S}}_2(\tilde{x}) \cdot \tilde{b} \cdot \tilde{p}(\tilde{x}_1, \dots, \tilde{x}_d) d\tilde{x}_{k+1} \dots d\tilde{x}_d = \tilde{a}^\top \cdot \left[\int\limits_{\mathbb{R}^{d-k}} \nabla^2 \tilde{p}(\tilde{x}) \cdot d\tilde{x}_{k+1} \cdots d\tilde{x}_d \right] \cdot \tilde{b}$$

$$= \sum_{\substack{1 \leqslant i \leqslant d \\ k+1 \leqslant j \leqslant d}} \tilde{a}_i \tilde{b}_j \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} \cdots \left[\int\limits_{-\infty}^{\infty} \frac{\partial^2 \tilde{P}(\tilde{x})}{\partial \tilde{x}_i \partial \tilde{x}_j} \cdot d\tilde{x}_j \right] \cdot \prod_{\substack{k+1 \leqslant t \leqslant d \\ t \neq j}} d\tilde{x}_t = 0,$$

because for any $j \in \{k+1, \dots, n\}$: $\int_{-\infty}^{\infty} \frac{\partial^2 \tilde{P}(\tilde{x})}{\partial \tilde{x}_i \partial \tilde{x}_j} \cdot d\tilde{x}_j = 0$ due to Assumption **(A3)**, which leads to the desired result. □

In order to combine several values of y, we will estimate the matrix $\mathcal{V}_2 = \mathbb{E}\Big([\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big)$, which is the expectation with respect to y of the square (in the matrix multiplication sense) of the conditional expectation from Lemma 4, as well as $\mathcal{V}_2' = \mathbb{E}\Big(y^2[\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big)$, and consider the k largest eigenvectors (we made the weak assumption of existence of matrices $\mathcal{V}_2$ and $\mathcal{V}_2'$, which is satisfied for majority of problems). From the lemma above, this matrix has a column space included in the e.d.r. subspace, thus, if it has rank k, we get exact recovery by an eigenvalue decomposition.

Effect of slicing. We now study the effect of slicing and show that in the population case, it is superior to the non-sliced version, as it recovers the true e.d.r. subspace in more situations.

Proposition 2. *Assume **(A1)** and **(A3)**. The matrix w may be recovered from the moment in Lemma 3 (up to right linear transform) if and only if $\mathbb{E}[\nabla^2 g(w^\top x)]$ has full rank. If this condition is satisfied, then SPHD also recovers w up to scale.*

Proof. The first statement is a consequence of the proof of Lemma 3. Moreover, using the Lowner-Heinz theorem about operator convexity [33]:

$$\mathcal{V}_2' = \mathbb{E}\big[\mathbb{E}(y\mathcal{S}_2(x)|y)^2\big] \succcurlyeq \big[\mathbb{E}[\mathbb{E}(y\mathcal{S}_2(x)|y)]\big]^2 = \big[\mathbb{E}(y\mathcal{S}_2(x))\big]^2,$$

showing that the new moment is dominating the PHD moment, thus implying that

$$\operatorname{rank}[\mathcal{V}_2'] \geqslant \operatorname{rank}\big[\mathbb{E}(y\mathcal{S}_2(x))\big].$$

Therefore, if $\operatorname{rank}\big[\mathbb{E}(y\mathcal{S}_2(x))\big] = k$, then $\operatorname{rank}[\mathcal{V}_2'] = k$ (note that there is not such a simple proof for $\mathcal{V}_2$). □

Elliptically symmetric distributions. When x is a standard Gaussian random variable, then $\mathcal{S}_2(x) = -I + xx^\top$, and thus $\mathbb{E}\Big([\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big) = \mathbb{E}\Big([I - \operatorname{Cov}(x|y)]^2\Big)$, and we recover the sliced average variance estimation (SAVE) method by Cook and Weisberg [8]. However, our method applies to all distributions (with known score functions).

2.3. Relationship between first and second order methods

All considered methods have their own failure modes. The simplest one: ADE works only in single-index model and has quite a simple working condition : $\mathbb{E}[g'(w^\top x)] \neq 0$. The sliced improvement (e.g., SADE) of this algorithm has

a better performance, however it still suffers from symmetric cases, when the inverse regression curve is partly degenerated. PHD+ can not work properly in linear models and symmetric cases. SPHD is stronger than PHD+ and potentially has the widest application area among four described methods. See summary in Table 1.

Our conditions rely on the full possible rank of certain expected covariance matrices. When the function g is selected randomly from all potential functions from $\mathbb{R}^k$ to $\mathbb{R}$, rank-deficiencies typically do not occur and it would be interesting to show that indeed they appear with probability zero for certain random function models.

method	main equation	score	sliced	single-index	multi-index
ADE	$\mathbb{E}(\mathcal{S}_1(x)y)$	first	no	$\mathbb{E}[g'(w^\top x)] \neq 0$	does not work
SADE	$\mathbb{E}_y\Big[\mathbb{E}\big(\mathcal{S}_1(x)\vert y\big)^{\otimes 2}\Big]$	first	yes	$\mathbb{E}_y\Big[\Vert\mathbb{E}\big(\mathcal{S}_1(x)\vert y\Vert^2\Big] > 0$	$\operatorname{rank}\mathbb{E}_y\Big[\mathbb{E}\big(\mathcal{S}_1(x)\vert y\big)^{\otimes 2}\Big] = k$
PHD+	$\mathbb{E}(\mathcal{S}_2(x)y)$	second	no	$\mathbb{E}[g''(w^\top x)] \neq 0$	$\operatorname{rank}\big[\mathbb{E}[\nabla^2 g(x)]\big] = k$
SPHD	$\mathbb{E}_y\Big[\mathbb{E}\Big((\mathcal{S}_2(x)\vert y)^2\Big)\Big]$	second	yes	$\mathbb{E}_y\Big[\operatorname{tr}\mathbb{E}\Big((\mathcal{S}_2(x)\vert y)^2\Big)\Big] > 0$	$\operatorname{rank}\mathbb{E}_y\Big[\mathbb{E}\Big((\mathcal{S}_2(x)\vert y)^2\Big)\Big] = k$

TABLE 1
Comparison of different methods using score functions.

3. Estimation from finite sample

In this section, we consider finite sample estimators for the moments we have defined in Section 2. Since our extensions are combinations of existing techniques (using score functions and slicing) our finite-sample estimators naturally rely on existing work [17, 39].

In this section, we assume that the score function is known. We consider learning the score function in Section 4.

3.1. Estimator and algorithm for SADE

Our goal is to provide an estimator for $\mathcal{V}_{1,\mathrm{cov}} = \mathbb{E}\Big[\mathbb{E}\big(\mathcal{S}_1(x)\vert y\big)\mathbb{E}\big(\mathcal{S}_1(x)\vert y\big)^\top\Big] = \operatorname{Cov}\Big[\mathbb{E}\big(\mathcal{S}_1(x)\vert y\big)\Big]$ given a finite sample (x_i, y_i), $i = 1, \dots, n$. A similar estimator for $\mathcal{V}'_{1,\mathrm{cov}}$ could be derived. In order to estimate $\mathcal{V}_{1,\mathrm{cov}} = \operatorname{Cov}\Big[\mathbb{E}\big(\mathcal{S}_1(x)\vert y\big)\Big]$, we will use the identity

$$\operatorname{Cov}\big[\mathcal{S}_1(x)\big] = \operatorname{Cov}\Big[\mathbb{E}\big(\mathcal{S}_1(x)\vert y\big)\Big] + \mathbb{E}\Big[\operatorname{Cov}\big(\mathcal{S}_1(x)\vert y\big)\Big].$$

We use the natural consistent estimator $\frac{1}{n}\sum_{i=1}^n \mathcal{S}_1(x_i)\mathcal{S}_1(x_i)^\top$ of $\operatorname{Cov}\big[\mathcal{S}_1(x)\big]$.

In order to obtain an estimator of $\mathbb{E}\Big[\operatorname{Cov}\big(\mathcal{S}_1(x)\vert y\big)\Big]$, we consider slicing the real numbers in H different *slices*, $I_1, \dots, I_H$, which are contiguous intervals

that form a partition of $\mathbb{R}$ (or of the range of all y_i, $i = 1, \dots, n$). We then compute an estimator of the conditional expectation $(\mathcal{S}_1)_h = \mathbb{E}(\mathcal{S}_1(x)|y \in I_h)$ with empirical averages: denoting $\hat{p}_h$ the empirical proportion of y_i, $i = 1, \dots, n$, that fall in the slice I_h (which is assumed to be strictly positive), we estimate $\mathbb{E}(\mathcal{S}_1(x)|y \in I_h)$ by

$$(\hat{\mathcal{S}}_1)_h = \frac{1}{n\hat{p}_h} \sum_{i=1}^{n} 1_{y_i \in I_h} \mathcal{S}_1(x_i).$$

We then estimate $\mathrm{Cov}\big(\mathcal{S}_1(x)|y \in I_h\big)$ by

$$(\hat{\mathcal{S}}_1)_{\mathrm{cov},h} = \frac{1}{n\hat{p}_h - 1} \sum_{i=1}^{n} 1_{y_i \in I_h} \big(\mathcal{S}_1(x_i) - (\hat{\mathcal{S}}_1)_h\big)\big(\mathcal{S}_1(x_i) - (\hat{\mathcal{S}}_1)_h\big)^\top.$$

Note that it is important here to normalize the covariance computation by $\frac{1}{n\hat{p}_h - 1}$ (usual unbiased normalization of the variance) and not $\frac{1}{n\hat{p}_h}$, to allow consistent estimation even when the number of elements per slice is small (e.g., equal to 2).

We finally use the following estimator of $\mathcal{V}_{1,\mathrm{cov}}$:

$$\hat{\mathcal{V}}_{1,\mathrm{cov}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}(x_i)\mathcal{S}(x_i)^\top - \sum_{h=1}^{H} \hat{p}_h \cdot (\hat{\mathcal{S}}_1)_{\mathrm{cov},h}.$$

The final SADE algorithm is thus the following:

- Divide the range of $y_1, \dots, y_n$ into H slices $I_1, \dots, I_H$. Let $\hat{p}_h > 0$ be the proportion of y_i, $i = 1, \dots, n$, that fall in slice I_h.
- For each slice I_h, compute the sample mean $(\hat{\mathcal{S}}_1)_h$ and covariance $(\hat{\mathcal{S}}_1)_{\mathrm{cov},h}$: $(\hat{\mathcal{S}}_1)_h = \frac{1}{n\hat{p}_h} \sum_{i=1}^{n} 1_{y_i \in I_h} \mathcal{S}_1(x_i)$ and $(\hat{\mathcal{S}}_1)_{\mathrm{cov},h} = \frac{1}{n\hat{p}_h - 1} \sum_{i=1}^{n} 1_{y_i \in I_h} \big(\mathcal{S}_1(x_i) - (\hat{\mathcal{S}}_1)_h\big)\big(\mathcal{S}_1(x_i) - (\hat{\mathcal{S}}_1)_h\big)^\top$.
- Compute $\hat{\mathcal{V}}_{1,\mathrm{cov}} = \frac{1}{n} \sum_{i=1}^{n} \mathcal{S}(x_i)\mathcal{S}(x_i)^\top - \sum_{h=1}^{H} \hat{p}_h \cdot (\hat{\mathcal{S}})_{\mathrm{cov},h}$.
- Find the k largest eigenvalues and let $\hat{w}_1, \dots, \hat{w}_k$ be eigenvectors in $\mathbb{R}^d$ corresponding to these eigenvalues.

Choice of slices. There are different ways to choose slices $I_1, \dots, I_H$:

- all slices have the same length, that is we choose the maximum and the minimum of $y_1, \dots, y_n$, and divide the range of y into H equal slices (for simplicity, we assume that n is a multiple of H),
- we can also use the distribution of y to ensure a balanced distribution of observations in each slice, and choose $I_h = (\hat{F}_y^{-1}((h-1)/H), \hat{F}_y^{-1}(h/H)]$, where $\hat{F}_y(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{y_i \leqslant t}$ is the empirical distribution function of y. If n is a multiple H, there are exactly $c = n/H$ observations per slice.

Later on in our experiments, we use the second way, where every slice has $c = n/H$ points, and our consistency result applies to this situation as well. Note that there is a variation of SADE, which uses standardized data. If we denote the standardized data as $\tilde{\mathcal{S}}_1(x_i) = \Big[\frac{1}{n}\sum_{j=1}^{n} \mathcal{S}_1(x_j)\mathcal{S}_1(x_j)^\top\Big]^{-1/2}\mathcal{S}_1(x_i)$ (Remark 5.3 in [21]).

Computational complexity. The first step of the algorithm requires $O(n)$ elementary operations, the second $O(nd^2)$ operations, the third $O(Hd^2)$ and the fourth $O(kd^2)$ operations. The overall dependence on dimension d is quadratic, while the dependence on the number of observations is linear in n, as common in moment-matching methods.

Estimating the number of components. Our estimation method does not depend on k, up to the last step where the first k largest eigenvectors are selected. A simple heuristic to select k, similar to the selection of the number of components in principal component analysis, would select the largest k such that the gap between the k-th and $(k+1)$-th eigenvalue is large enough. This could be made more formal using the technique of [21] for sliced inverse regression.

3.2. Estimator and algorithm for SPHD

We follow the same approach as for the SADE algorithm above, leading to the following algorithm, which estimates $\mathcal{V}_2 = \mathbb{E}\Big([\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big)$ and computes its principal eigenvectors. Note that $\mathbb{E}\big[\mathbb{E}(\mathcal{S}_2(x)|y \in I_h)\big] = \mathbb{E}\big[\mathcal{S}_2(x)\big] = 0$.

- Divide the range of $y_1, \dots, y_n$ into H slices $I_1, \dots, I_H$. Let $\hat{p}_h > 0$ be the proportion of y_i, $i = 1, \dots, n$, that fall in slice I_h.
- For each slice, compute the sample mean $(\hat{\mathcal{S}}_2)_h$ of $\mathcal{S}_2(x)$: $(\hat{\mathcal{S}}_2)_h = \dfrac{1}{n\hat{p}_h}\displaystyle\sum_{i=1}^{n} 1_{y_i \in I_h}\mathcal{S}_2(x_i)$.
- Compute the weighted covariance matrix $\hat{\mathcal{V}}_2 = \displaystyle\sum_{h=1}^{H} \hat{p}_h(\hat{\mathcal{S}}_2)_h^2$, find the k largest eigenvalues and let $\hat{w}_1, \dots, \hat{w}_k$ be eigenvectors corresponding to these eigenvalues.

The matrix $\hat{\mathcal{V}}_2$ is then an estimator of $\mathbb{E}\Big([\mathbb{E}\big(\mathcal{S}_2(x)|y\big)]^2\Big)$.

Computational complexity. The first step of the algorithm requires $O(n)$ elementary operations, the second $O(nd^2)$ operations, the third $O(Hd^3)$ and the fourth $O(kd^2)$ operations. The overall dependence on dimension d is cubic, hence the method is slower than SADE (but still linear in the number of observations n).

3.3. Consistency for the SADE estimator and algorithm

In this section, we prove the consistency of the SADE moment estimator and the resulting algorithm, when the score function is known. Following Hsing and Carroll [17] and Xhu and Ng [39], we can get $\sqrt{n}$-consistency for the SADE algorithm with very broad assumptions regarding the problem.

In this section, we focus on the simplest set of assumptions to pave the way to the analysis for the nuclear norm in future work. The key novelty compared to [17, 39] is a precise *non-asymptotic* analysis with precise constants.

We make the following assumptions:

(L1) The function $m : \mathbb{R} \to \mathbb{R}^d$ such that $\mathbb{E}(\ell(x)|y) = m(y)$ is L-Lipschitz-continuous.

(L2) The random variable $y \in \mathbb{R}$ is sub-Gaussian, i.e., such that $\mathbb{E}e^{t(y-Ey)} \leqslant e^{\tau_y^2 t^2/2}$, for some $\tau_y > 0$.

(L3) The random variables $\ell_j(x) \in \mathbb{R}$ are sub-Gaussian, i.e., such that $\mathbb{E}e^{t\ell_j(x)} \leqslant e^{\tau_\ell^2 t^2/2}$ for each component $j \in \{1, \dots, d\}$, for some $\tau_\ell > 0$.

(L4) The random variables $\eta_j = \ell_j(x) - m_j(y) \in \mathbb{R}$ are sub-Gaussian, i.e., such that $\mathbb{E}e^{t\eta_j} \leqslant e^{\tau_\eta^2 t^2/2}$ for each component $j \in \{1, \dots, d\}$, for some $\tau_\eta > 0$.

Now we formulate and proof the main theorem, where $\|\cdot\|_*$ is the nuclear norm, defined as $\|A\|_* = \operatorname{tr}\big(\sqrt{A^T A}\big)$:

Theorem 1. *Under assumptions **(L1)** - **(L4)** we get the following bound on $\|\hat{\mathcal{V}}_{1,cov} - \mathcal{V}_{1,cov}\|_*$: for any $\delta < \dfrac{1}{n}$, with probability not less than $1 - \delta$:*

$$\|\hat{\mathcal{V}}_{1,cov} - \mathcal{V}_{1,cov}\|_* \leqslant \frac{d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2)}{\sqrt{n}} \sqrt{\log \frac{24d^2}{\delta}} + \frac{8L^2\tau_y^2 + 16\tau_\eta\tau_y L\sqrt{d} + (157\tau_\eta^2 + 2\tau_\ell^2)d\sqrt{d}}{n} \log^2 \frac{32d^2 n}{\delta}.$$

The proof of the theorem can be found in Appendix A.2. The leading term is proportional to $\dfrac{d\sqrt{d}\tau_\eta^2}{\sqrt{n}}$, with a tail which is sub-Gaussian. We thus get a $\sqrt{n}$-consistent estimator or $\mathcal{V}_1$. The dependency in d could probably be improved, in particular when using slices of sizes c that tend to infinity (as done in [24]).

4. Learning score functions

All previous methods can work only if we know the score function of first or second order. In practice, we do not have such information, and we have to learn score functions from sampled data. In this section, we only consider the first-order score function $\ell(x) = \mathcal{S}_1(x) = -\nabla \log p(x)$.

We first present the score matching approach of [1], and then apply it to our problem.

4.1. Score matching to estimate score from data

Given the true score function $\ell(x) = \mathcal{S}_1(x) = -\nabla \log p(x)$ and some i.i.d. data generated from $p(x)$, score matching aims at estimating the parameter of a model for the score function $\hat{\ell}(x)$, by minimizing a empirical quantity aiming to estimate

$$\mathcal{R}_{\text{score}}(\hat{\ell}) = \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\ell(x) - \hat{\ell}(x)\|^2 dx.$$

As is, the quantity above leads to consistent estimation (i.e., pushing $\hat{\ell}$ close to ℓ), but seems to need the knowledge of the true score $\ell(x)$. A key insight from [1] is to use integration by parts to get (assuming the integrals exist):

$$\begin{aligned}
\mathcal{R}_{\text{score}}(\hat{\ell}) &= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\big[\|\ell(x)\|^2 + \|\hat{\ell}(x)\|^2 + 2\hat{\ell}(x)^\top \nabla \log p(x)\big] dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\ell(x)\|^2 dx + \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\hat{\ell}(x)\|^2 dx + \int_{\mathbb{R}^d} \hat{\ell}(x)^\top \nabla p(x) dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\ell(x)\|^2 dx + \frac{1}{2}\int_{\mathbb{R}^d} p(x)\|\hat{\ell}(x)\|^2 dx - \int_{\mathbb{R}^d} (\nabla \cdot \hat{\ell})(x) p(x) dx,
\end{aligned}$$

by integration by parts, where $(\nabla \cdot \hat{\ell})(x) = \sum_{i=1}^d \frac{\partial \hat{\ell}}{\partial x_i}(x)$ is the divergence of $\hat{\ell}$ at x [2].

The first part of the last right hand side does not depend on $\hat{\ell}$ while the two other parts are expectations under $p(x)$ of quantities that only depend on $\hat{\ell}$. Thus is can we well approximated, up to a constant, by:

$$\hat{\mathcal{R}}_{\text{score}}(\hat{\ell}) = \frac{1}{2n}\sum_{i=1}^n \|\hat{\ell}(x_i)\|^2 - \frac{1}{n}\sum_{i=1}^n (\nabla \cdot \hat{\ell})(x_i).$$

Parametric assumption. If we assume that the score is linear combination of finitely many basis functions, we will get a consistent estimator of these parameters. That is, we make the following assumption:

(A4) The score function $\ell(x)$ is a linear combination of known basis functions $\psi^j(x)$, $j = 1, \dots, m$, where $\psi^j : \mathbb{R}^d \to \mathbb{R}^d$, that is $\ell(x) = \sum_{j=1}^m \psi^j(x)\theta_j^*$, for some $\theta^* \in \mathbb{R}^m$. We assume that the score function and its derivatives are squared-integrable with respect to $p(x)$.

In this paper, we consider for simplicity a parametric assumption for the score. In order to go towards non-parametric estimation, we would need to let the number m of basis functions to grow with n (exponentially with no added assumptions), and this is an interesting avenue for future work. In simulations in Section 5, we consider a simple set of basis function which are localized functions around observations; these can approximate reasonably well in practice most densities and led to good estimation of the e.d.r. subspace. Moreover, if we have the additional knowledge that the components of x are statistically

independent (potentially after linearly transforming them using independent component analysis [18]), we can use separable functions for the scores.

We introduce the notation

$$\Psi(x) = \begin{pmatrix} \psi_1^1(x) & \cdots & \psi_d^1(x) \\ \vdots & \ddots & \vdots \\ \psi_1^m(x) & \cdots & \psi_d^m(x) \end{pmatrix} \in \mathbb{R}^{m \times d},$$

so that the score function $\hat{\ell}$ we wish to estimate has the form

$$\hat{\ell}(x) = \Psi(x)^\top \theta \in \mathbb{R}^d.$$

We also introduce the notation $(\nabla \cdot \Psi)(x) = \begin{pmatrix} (\nabla \cdot \Psi^1)(x) \\ \vdots \\ (\nabla \cdot \Psi^m)(x) \end{pmatrix} \in \mathbb{R}^m$, so that $(\nabla \cdot \hat{\ell})(x) = \theta^\top (\nabla \cdot \Psi)(x)$.

The empirical score function may then be written as:

$$\hat{\mathcal{R}}_{\text{score}}(\theta) = \frac{1}{2}\theta^\top \Big(\frac{1}{n}\sum_{i=1}^n \Psi(x_i)\Psi(x_i)^\top\Big)\theta - \theta^\top \Big(\frac{1}{n}\sum_{i=1}^n (\nabla \cdot \Psi)(x_i)\Big), \tag{4.1}$$

which is a quadratic function of θ and can thus be minimized by solving a linear system in running time $O(m^3 + m^2 dn)$ (to form the matrix and to solve the system).

Given standard results regarding the convergence of $\frac{1}{n}\sum_{i=1}^n \Psi(x_i)\Psi(x_i)^\top \in \mathbb{R}^{m \times m}$ to its expectation and of $\frac{1}{n}\sum_{i=1}^n (\nabla \cdot \hat{\ell})(x_i) \in \mathbb{R}^m$ to its expectation, we get a $\sqrt{n}$-consistent estimation of θ^* under simple assumptions (see Theorem 2).

4.2. Score matching for sliced inverse regression: two-step approach

We can now combine our linear parametrization of the score with the SIR approach outlined in Section 3. The true conditional expectation is

$$\mathbb{E}\big(\ell(x)|y\big) = \sum_{j=1}^m \mathbb{E}\big(\psi^j(x)|y\big)\theta_j^*,$$

and belongs to the e.d.r. subspace. We consider H different slices $I_1, \dots, I_H$, and the following estimator, which simply replaces the true score by the estimated score (i.e., θ^* by θ), and highlights the dependence in θ.

The estimator $\hat{\mathcal{V}}_{1,\text{cov}}$ can be rewritten as

$$\hat{\mathcal{V}}_{1,\text{cov}} = \sum_{i=1}^n \sum_{j=1}^n \frac{\alpha_{i,j}}{n(|I_h(i,j)| - 1)} \ell(x_i)\ell(x_j)^\top,$$

where

$$\alpha_{i,j} = \begin{cases} 1 & \text{if } i \neq j \text{ in the same slice} \\ 0 & \text{otherwise} \end{cases}$$

Using linear property $\ell(x) = \Psi(x)^\top \theta$:

$$\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\alpha_{i,j}}{|I_h(i,j)|-1}\Psi(x_i)^\top\theta\theta^\top\Psi(x_i). \tag{4.2}$$

In the two-step approach, we first solve the score matching optimization problem to obtain an estimate for the optimal parameters θ^* and then use them to get the k largest eigenvectors of covariance matrix $\hat{\mathcal{V}}_1$. This approach works well in low dimensions when the score function can be well approximated; otherwise, we may suffer from the curse of dimensionality: if we want a good approximation of the score function, we would need an exponential number of basis functions. In Section 4.3, we consider a direct approach aiming at improving robustness.

Consistency. Let also provide the result of consistency in the case of unknown score function under Assumption **(A4)** for the two-step algorithm. We will make the additional assumptions:

(M1) The random variables $\Psi_i^j(x) \in \mathbb{R}$ are sub-Gaussian, i.e., such that $\mathbb{E}e^{t(\Psi_i^j(x) - E\Psi_i^j(x))} \leqslant e^{\tau_\Psi^2 t^2/2}$, for some $\tau_\Psi > 0$.

(M2) The random variables $(\nabla\cdot\Psi^i)(x)$ are sub-Gaussian, i.e., such that $\mathbb{E}e^{t((\nabla\cdot\Psi^i)(x) - E(\nabla\cdot\Psi^i)(x)} \leqslant e^{\tau_{\nabla\Psi}^2 t^2/2}$, for some $\tau_{\nabla\Psi} > 0$.

(M3) The matrix $\mathbb{E}\big[\Psi(x)\Psi(x)^\top\big]$ is not degenerated and we let λ_0 denote its minimal eigenvalue.

Theorem 2. *Let $\hat{\theta}$ be the estimated θ, obtained on the first step of algorithm. Under Assumptions **(A1)**, **(A2)**, **(A4)**, **(L1)** - **(L4)** and **(M1)**, **(M2)**, **(M3)**, for $\delta \leqslant 1/n$ and n large enough, that is, $n > c_1 + c_2 \log\frac{1}{\delta}$ for some positive constants c_1 and c_2 not depending on n and δ:*

$$\|\mathcal{V}_{1,cov}(\theta^*) - \hat{\mathcal{V}}_{1,cov}(\hat{\theta})\|_* \leqslant \frac{1}{\sqrt{n}}\sqrt{2\log\frac{48m^2d}{\delta}}\times$$

$$\left[d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2) + \frac{9m}{2}\mathbb{E}\|\Psi(x)\|_2^2\cdot\|\theta^*\|\left(\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\lambda_0} + \frac{4\|b\|\tau_\Psi^2\sqrt{m^3d}}{\lambda_0}\right)\right] + O\left(\frac{1}{n}\right).$$

with probability not less than $1-\delta$.

Now, let us formulate and proof non-asymptotic result about the real and the estimated e.d.r. spaces. We need to define a notion of distance between subspaces spanned by two sets of vectors. We use the square trace error $R^2(w, \hat{w})$ [15]:

$$R^2(w,\hat{w}) = 1 - \frac{1}{k}\mathrm{tr}\big[w\cdot\hat{w}\big], \tag{4.3}$$

where $w \in \mathbb{R}^{k\times d}$ and $\hat{w} \in \mathbb{R}^{k\times d}$ both have orthonormal columns. It is always between zero and one, and equal to zero if and only if $w = \hat{w}$. This distance is closely related to *principal angles* notion:

$$\sin\Theta(\hat{w}^\top w) = \operatorname{diag}(\sin(\cos^{-1}\sigma_1), \dots, \sin(\cos^{-1}\sigma_d),$$

where $\sigma_1, \dots, \sigma_d$ are the singular values of $\hat{w}^\top w$. Actually, $R(w, \hat{w}) \cdot \sqrt{k} = \|\sin\Theta(\hat{w}^\top w)\|_F$.

We use Davis-Kahan "$\sin\theta$ theorem" [27, Theorem V.3.6] in the following form [37, Theorem 2]:

Theorem 3. *Let Σ and $\hat{\Sigma} \in \mathbb{R}^{d\times d}$ be symmetric, with eigenvalues $\lambda_1 \geqslant \dots \geqslant \lambda_d$ and $\hat{\lambda}_1, \dots, \hat{\lambda}_d$ respectively. Fix $1 \leqslant r \leqslant s \leqslant d$, let $k = s - r + 1$, and let $U = (u_r, u_{r+1}, \dots, u_s) \in \mathbb{R}^{d\times k}$ and $\hat{U} = (\hat{u}_r, \hat{u}_{r+1}, \dots, \hat{u}_s) \in \mathbb{R}^{d\times k}$ have orthonormal columns satisfying $\Sigma u_j = \lambda_j u_j$ and $\hat{\Sigma}\hat{u}_j = \hat{\lambda}_j \hat{u}_j$ for $j = r, \dots, s$. Let $\delta = \inf\{|\hat{\lambda} - \lambda| : \lambda \in [\lambda_s, \lambda_r],\ \hat{\lambda} \in (-\infty, \hat{\lambda}_{s+1}] \cup [\hat{\lambda}_{r-1}, +\infty)$, where $\hat{\lambda}_0 = +\infty$ and $\hat{\lambda}_{p+1} = -\infty$. Assume, that $\delta > 0$, then:*

$$R(U, \hat{U}) \leqslant \frac{\|\hat{\Sigma} - \Sigma\|_F}{\delta\sqrt{k}}. \tag{4.4}$$

Using Corollary 4.1 and its discussion in [30], we can derive, that $R(w, \hat{w}) \leqslant \dfrac{\|\mathcal{V}_1(\theta^*) - \hat{\mathcal{V}}_1(\hat{\theta})\|_F}{|\lambda_k - \hat{\lambda}_{k+1}|\sqrt{k}}$, where w and $\hat{w}$ are the real and the estimated e.d.r. spaces respectively. Using Weyl's inequality [27] : if $\|\mathcal{V}_1(\theta^*) - \hat{\mathcal{V}}_1(\hat{\theta})\|_2 < \varepsilon \ \Rightarrow |\hat{\lambda}_i - \lambda_i| < \varepsilon$ for all $i = 1, \dots, d$ and taking $\varepsilon = (\lambda_k - \lambda_{k+1})/2 = \lambda_k/2$ we get:

$$R(w, \hat{w}) \leqslant \frac{2\|\mathcal{V}_1(\theta^*) - \hat{\mathcal{V}}_1(\hat{\theta})\|_*}{\lambda_k\sqrt{k}}, \text{ if } \|\mathcal{V}_1(\theta^*) - \hat{\mathcal{V}}_1(\hat{\theta})\|_F < \lambda_k/2,$$

because $\|\cdot\|_F \leqslant \|\cdot\|_*$. We can now formulate our main theorem for the analysis of the SADE algorithm:

Theorem 4. *Consider assumptions **(A1)**, **(A2)**, **(A4)**, **(L1)**- **(L4)**, **(M1)**, **(M2)**, **(M3)** and:*

(K1) *The matrix $\mathcal{V}_1$ has a rank k: $\lambda_k > 0$.*

For $\delta \leqslant 1/n$ and n large enough: $n > c_1 + c_2 \log\frac{1}{\delta}$ for some positive constants c_1 and c_2 not depending on n and δ:

$$R(\hat{w}, w) \leqslant \frac{2}{\sqrt{n}\lambda_k\sqrt{k}}\sqrt{2\log\frac{48m^2d}{\delta}} \times$$

$$\left[d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2) + \frac{9m}{2}\mathbb{E}\|\Psi(x)\|_2^2 \cdot \|\theta^*\|\left(\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\lambda_0} + \frac{4\|b\|\tau_\Psi^2\sqrt{m^3d}}{\lambda_0}\right)\right] + O\left(\frac{1}{n}\right).$$

with probability not less than $1 - \delta$.

We thus obtain a convergence rate in $1/\sqrt{n}$, with an explicit dependence on all constants of the problem. The dependence on dimension d and number of basis functions m is of the order (forgetting logarithmic terms): $O(\frac{d^{3/2}}{n^{1/2}} + \frac{m^{5/2}}{n^{1/2}}\|\theta^*\|)$. For $k > 1$, our dependence on the sample size n is improved compared to existing work such as [11] (while it matches the dependence for $k = 1$ with [16]), but this comes at the expense of assuming that the score functions satisfy a parametric model.

For a fixed number of basis functions m, we thus escape the curse of dimensionality as we get a polynomial dependence on d (which we believe could be improved). However, when no assumptions are made on score functions, the number m and potentially the norm $\|\theta^*\|$ has to grow with the number of observations n. We are indeed faced with a traditional non-parametric estiamation problem, and the number m will need to grow when n grows depending on the smoothness assumptions we are willing to make on the score function, with an effect on θ^* (and probably λ_0, which we will neglect in the discussion below). While a precise analysis is out of the scope of the paper and left for future work, we can make an informal argument as follows: in order to approximate the score with precision ε with a set of basis function where $\|\theta^*\|$ is bounded, we need $m(\varepsilon)$ basis functions. Thus, we end up with two sources of errors, an approximation error ε and an estimation error of order $O(m(\varepsilon)^{5/2}/n^{1/2})$. Typically, if we assume that the score has a number of bounded derivatives proportional to dimension or if we assume the input variables are independent and we can thus estimate the score *independently* for each dimension, $m(\varepsilon)$ is of the order $1/\varepsilon^{1/r}$, where r is independent of the dimension, leading to an overall rate which is independent of the dimension.

4.3. Score matching for SIR: direct approach

We can also try to combine these two steps to try to avoid the "curse of dimensionality". Our estimation of the score, i.e., of the parameter θ is done only to be used within the SIR approach where we expect the matrix $\hat{\mathcal{V}}_{1,\mathrm{cov}}$ to have rank k. Thus when estimating θ by minimizing $\hat{\mathcal{R}}_{\mathrm{score}}(\theta)$, we may add a regularization that penalizes large ranks for $\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta) = \frac{1}{n}\sum_{i,j=1}^{n} \frac{\alpha_{i,j}}{|I_h(i,j)|-1}\Psi(x_i)^\top\theta\theta^\top\Psi(x_i)$, where we highlight the dependence on $\theta \in \mathbb{R}^m$. By enforcing the low-rank constraint, our aim is to circumvent a potential poor estimation of the score function, which could be enough for the task of estimating the e.d.r. space (we see a better behavior in our simulations in Section 5).

Introduce matrix $\mathcal{L}(\theta) = (\Psi(x_i)^\top\theta, \dots, \Psi(x_n)^\top\theta) \in \mathbb{R}^{d\times n}$ and $A \in \mathbb{R}^{n\times n}$ with $A_{i,j} = \frac{\alpha_{i,j}}{n\cdot(|I_h(i,j)|-1)}$ and $\mathcal{A}(\theta) = \mathcal{L}\cdot A^{1/2} \in \mathbb{R}^{d\times n}$.

We may then penalize the nuclear norm of $\mathcal{A}(\theta)$, or potentially consider norms that take into account that we look for a rank k (e.g., the k-support norm on the spectrum of $\mathcal{A}(\theta)$ [25]). We have,

$$\|\mathcal{A}(\theta)\|_* = \mathrm{tr}(\mathcal{A}(\theta)\mathcal{A}(\theta)^\top)^{1/2} = \mathrm{tr}\big[\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta)^{1/2}\big].$$

Combining two penalties, we have a convex optimization task:

$$\hat{\mathcal{R}}(\theta) = \hat{\mathcal{R}}_{\text{score}}(\theta) + \lambda \cdot \operatorname{tr}\big[\hat{\mathcal{V}}_{1,\text{cov}}(\theta)^{1/2}\big]. \tag{4.5}$$

Efficient algorithm. Following [3], we consider reweighted least-squares algorithms. The trace norm admits the variational form (see [4]):

$$\|W\|_* = \frac{1}{2} \inf_{D \succ 0} \operatorname{tr}(W^T D^{-1} W + D).$$

The optimization problem (4.5) can be reformulated in the following way:

$$\theta \leftarrow \operatorname*{argmin}_{\theta} \ \hat{\mathcal{R}}_{\text{score}}(\theta) + \frac{\lambda}{2} \operatorname{tr}\Big(\mathcal{A}(\theta)^\top D^{-1} \mathcal{A}(\theta)\Big) \ \text{ and}$$

$$D \leftarrow \big(\mathcal{A}(\theta)\mathcal{A}(\theta)^\top + \varepsilon I_d\big)^{1/2}.$$

Note that the objective function is a quadratic function of θ. Decompose matrix $\mathcal{A}$ in the form:

$$\mathcal{A}(\theta) = \sum_{k=1}^{m} \theta_k \mathcal{A}_k, \quad \text{where } \mathcal{A}_k \in \mathbb{R}^{d \times n}.$$

Rearrange the regularizer term:

$$\operatorname{tr}\Big(\mathcal{A}(\theta)^\top D^{-1} \mathcal{A}(\theta)\Big) = \sum_{k=1}^{m} \sum_{l=1}^{m} \operatorname{tr}\Big[\theta_k \mathcal{A}_k^\top D^{-1} \mathcal{A}_l \theta_l\Big] = \theta^\top \mathcal{Y} \theta,$$

where

$$\mathcal{Y}_{k,l} = \operatorname{tr}\big[\mathcal{A}_k^\top D^{-1} \mathcal{A}_l\big].$$

Introduce notation:

$$\tilde{\mathcal{A}} = \Big(\operatorname{vect}(\mathcal{A}_1), \dots, \operatorname{vect}(\mathcal{A}_m)\Big) \in \mathbb{R}^{dn \times m}$$

$$\tilde{\mathcal{B}} = \Big(\operatorname{vect}(D^{-1}\mathcal{A}_1), \dots, \operatorname{vect}(D^{-1}\mathcal{A}_m)\Big) \in \mathbb{R}^{dn \times m},$$

then

$$\mathcal{Y} = \tilde{\mathcal{A}}^\top \tilde{\mathcal{B}}.$$

We can now estimate the complexity of this algorithm. Firstly we need to evaluate the quadratic form in $\hat{\mathcal{R}}_{\text{score}}(\theta)$: it is a summation of n multiplications of $m \times d$ and $d \times m$ matrices. Complexity of this step is $O(nm^2d)$. Secondly, we need to evaluate matrix matrices D^{-1}, $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{B}}$ using $O(d^3)$, $O(dHm)$ and $O(m \times d^2 H)$ operations respectively. Next, we need to evaluate matrix $\mathcal{Y}$, using $O(dnm^2)$ operations. Finally, evaluating $\mathcal{A}(\theta)$ requires $O(dnm)$ operations, $\mathcal{A}(\theta)\mathcal{A}(\theta)^\top$ requires $O(d^2 n)$ operations and evaluation of D requires $O(d^3)$ operations. Combining complexities, we get $O(md^2n + d^3 + nm^2d)$ operations, which is still linear in n.

5. Experiments

In this section we provide numerical experiments for SADE, PHD+ and SPHD on different functions f. We denote the true and estimated e.d.r. subspaces as $\mathcal{E}$ and $\hat{\mathcal{E}}$ respectively, defined from w and $\hat{w}$.

5.1. Known score functions

Consider a Gaussian mixture model with 2 components in $\mathbb{R}^d$:

$$p(x) = \sum_{i=1}^{2} \theta_i \cdot \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_i|^{1/2}} \cdot \exp\left\{ -\frac{1}{2}(X-\mu_i)^\top \Sigma_i^{-1}(X-\mu_i) \right\}, \quad (5.1)$$

where $\theta = (6/10, 4/10)$, $\mu_1 = (\underbrace{-1, \dots, -1}_{d})$, $\mu_2 = (\underbrace{1, \dots, 1}_{d})$, $\Sigma_1 = I_d$, $\Sigma_2 = 2 \cdot I_d$.
Contour lines of this distribution, when $d = 2$ are shown in Figure 1.

The error ε has a standard normal distribution. To estimate the effectiveness of an estimated e.d.r. subspace, we use the square trace error $R^2(w, \hat{w})$

$$R^2(w, \hat{w}) = 1 - \frac{1}{k}\mathrm{tr}\big[P \cdot \hat{P}\big],$$

where w and $\hat{w}$ are the real and the estimated e.d.r. vectors respectively and P and $\hat{P}$ are projectors, corresponding to these matrices. Note, that (4.3) is the special case of this formula with orthonormal matrices.

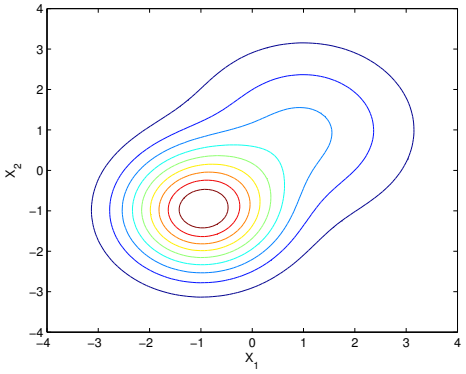

FIG 1. *Contour lines of Gaussian mixture pdf in 2D case.*

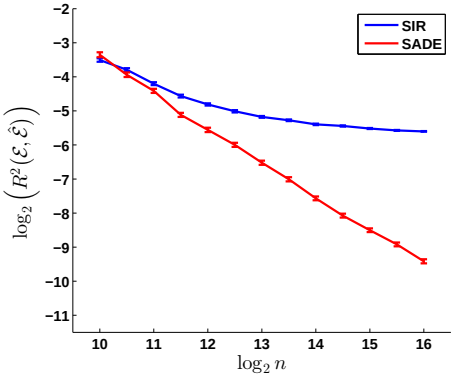

FIG 2. *Mean and standard deviation of* $R^2(\mathcal{E}, \hat{\mathcal{E}})$ *divided by* 10 *for the rational function (5.2).*

To show the dominance of SADE over SIR (which should only work for elliptically symmetric distributions), we consider the rational multi-index model

of the form

$$d = 10; \quad y = \frac{x_1}{1/2 + (x_2 + 2)^2} + \sigma \cdot \varepsilon. \tag{5.2}$$

Here the real e.d.r. space is generated by 2 specific vectors: $(1, 0, 0, \ldots, 0)$ and $(0, 1, 0, \ldots, 0)$. The number of slices is $H = 10$, and we consider numbers of observations n from 2^{10} to 2^{16} equally spaced in logarithmic scale, and we conduct 100 replicates to obtain means and standard deviations of the logarithms of square trace errors divided by $\sqrt{100} = 10$ (to assess significance of differences) as shown in Figure 2. Even in this simple model, the ordinary SIR algorithm does not work properly, because the distribution of the inputs x has no elliptical symmetry. When $n \to \infty$, the squared trace error tends to some nonzero constant depending on the properties of the density function, whereas SADE shows good performance with slope -1 (corresponding to a $\sqrt{n}$-consistent estimator).

Now, we compare the moments methods SADE, PHD+, SPHD. Although the goal of this paper is to compare moment matching techniques, we compare them with the state-of-the-art MAVE method [34], [31], [32]. It is worth noting two properties which we have already discussed concerning these methods:

1. Sliced methods have a wider application area than unsliced ones: SADE is stronger than ADE and SPHD is stronger than PHD+.
2. SADE can not recover the entire e.d.r. space in several cases, for example, a classification task or symmetric cases. For those cases, we should use second-order methods (i.e., methods based on $\mathcal{S}_2$).
3. MAVE works better, but the goal of this paper is to compare moment-matching techniques. In Figure 11, we provide examples where MAVE suffers from the curse of dimensionality and performs worse that moment-matching methods.

We conduct 3 experiments, where $H = 10$, $d = 10$, the error term ε has a normal standard distribution; numbers of observation n from 2^{10} to 2^{16} equally spaced in logarithmic scale and we made 10 replicas to evaluate sample means and variations:

- Rational model of the form:

$$y = \sum_{i=1}^{k} \tanh(8x_i - 16) \cdot i + \tanh(8x_i + 16) \cdot (k + 1 - i) + \varepsilon/4, \tag{5.3}$$

  where the effective reduction subspace dimension is $k = 2$.
  Results are shown of Figure 3 and we can see, that first-order method SADE works better, than second-order SPHD + and SPHD works better, than PHD+, that is slicing make the method more robust.
- Classification problem of the form

$$y = 1_{x_1^2 + 2x_2^2 > 4} + \varepsilon/4. \tag{5.4}$$

  We can see, that the error of SADE is close to 0.5 (Figure 4). This means that the method finds only one direction in the e.d.r. space. Moreover, SPHD gave better results than PHD+.

- Quadratic model of the form

$$d = 10; \quad y = x_1(x_1 + x_2 - 3) + \varepsilon. \tag{5.5}$$

Both SADE and SPHD show a good performance (Figure 5), while PHD+ can not recover the desired projection due to linearity of the function g.

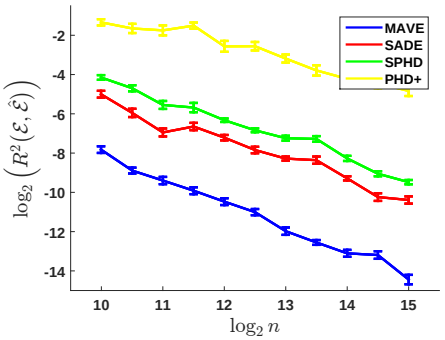

FIG 3. *Mean and standard deviation of* $R^2(\mathcal{E},\hat{\mathcal{E}})$ *divided by* $\sqrt{10}$ *for the rational function (5.3) with* $\sigma = 1/4$.

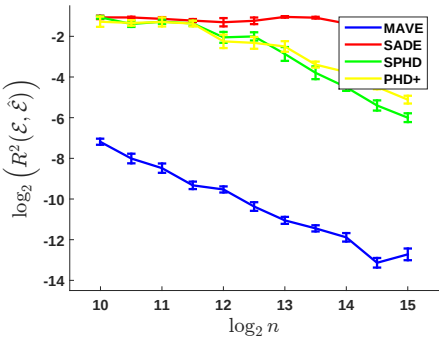

FIG 4. *Mean and standard deviation of* $R^2(\mathcal{E},\hat{\mathcal{E}})$ *divided by* $\sqrt{10}$ *for the the classification problem (5.4).*

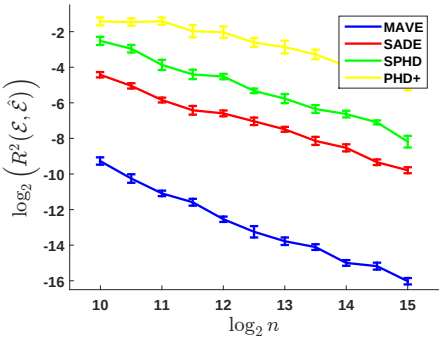

FIG 5. *Mean and standard deviation of* $R^2(\mathcal{E},\hat{\mathcal{E}})$ *divided by* $\sqrt{10}$ *for the quadratic (5.2) with* $\sigma = 1/4$.

5.2. *Unknown score functions*

Now, we conduct numerical experiments for SADE with unknown score functions. We consider a "toy" experiment and first examine results of the 2-step algorithm. We consider again a Gaussian mixture model (5.1) with 2 components in $\mathbb{R}^2$, where $\theta = (0.6, 0.4)$, $\mu_1 = (-0.5, 0.5), \mu_2 = (0.5, 0.5)$, $\Sigma_1 = 0.3 \cdot I_2$, $\Sigma_2 = 0.4 \cdot I_2$, with $y = \sin(x_1 + x_2) + \varepsilon$, where error ε has a standard normal

distribution. We choose these parameters of Gaussian mixture to make sure, that the probability of the vector X to be in the square $[-2,2]^2$ is close to 1.

We chose 100 Gaussian kernels as basis functions:

$$\psi_{i,j}(x) = \nabla \exp\left\{ -\frac{\|X - X_{i,j}\|^2}{2h^2} \right\}, 1 \leqslant i,j \leqslant 10,$$

where $X_{i,j}$ form the uniform grid on the square $[-2,2]^2$ (note that this does not imply any notion of Gaussianity for the underlying density, and the Gaussian kernel here could be replaced by any differentiable local function).

In practice, $\hat{\Psi}$ in (4.1) is close to be degenerated, and we use a regularized estimator $\theta^* = -\left(\hat{\Psi} + \alpha I_T\right)^{-1} \cdot \hat{\Phi}$ instead.

We choose $\alpha = 0.01$, $\sigma = 1$ and conduct 100 replicates with numbers of observations n from 2^{10} to 2^{16} equally spaced in logarithmic scale. The results of the experiments are presented in Figure 6: the square trace error tends to zero as n increases.

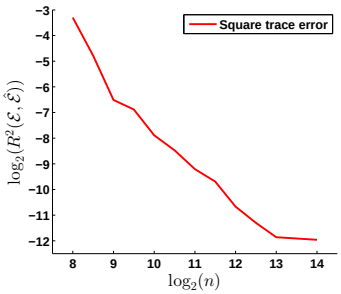

FIG 6. *Square trace error $R(\mathcal{E},\hat{\mathcal{E}})$ for different sample sizes.*

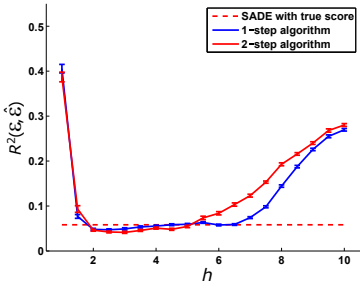

FIG 7. *Quadratic model, $d = 10$, $n = 1000$.*

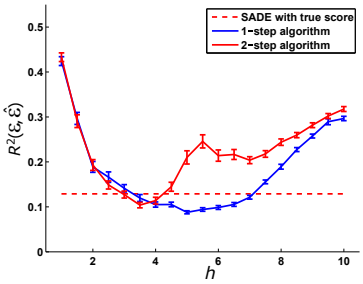

FIG 8. *Rational model, $d = 10$, $n = 1000$.*

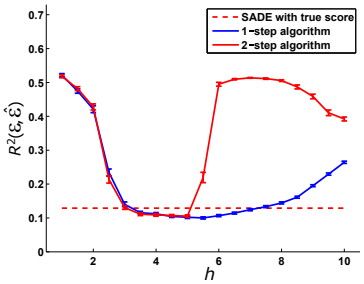

FIG 9. *Rational model, $d = 20$, $n = 2000$.*

In high dimensions, we can not use a uniform grid because of the curse of dimensionality. Instead of this, we will use n Gaussian kernels, centered in the

sample points X_i. Note here that we can only recover an approximate score function, but our goal is the estimation of the e.d.r. subspace.

We use our reweighted least-squares algorithm to solve the convex problem (4.5). We conduct several experiments on functions from the previous section.

We plot the performance as a function of the kernel bandwidth h to assess the robustness of the methods. On Figure 7 we provide the relationship between the square trace error and h for quadratic model (5.5) for $d = 10$ and $n = 1000$. In Figures 8, 9, 10, we provide the relationship between the square trace error and σ for rational model (5.2) for $d = 10$, $n = 1000$; $d = 20, n = 2000$ and $d = 50, n = 5000$. We see that for large d, the one-step algorithm is more robust than the two-step algorithm. Moreover, the experiments show that even with weak score functions, correct estimation of the e.d.r. space can be performed.

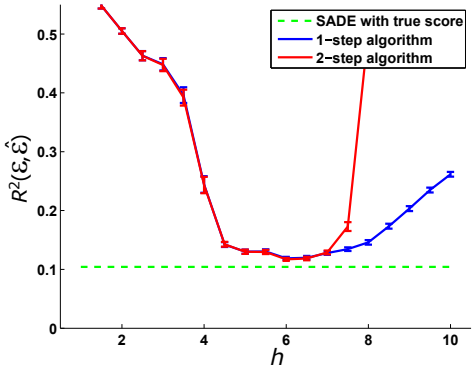

FIG 10. *Rational model,* $d = 50$, $n = 5000$.

Comparison with MAVE. We consider the rational model (5.3) with $n = 10000$, $k = 10$ and dimension d of data in the range $[20, 150]$. Probability density $p(x)$ has independent components, each of which has a form of mixture of 2 Gaussians with weights $(6/10, 4/10)$, means $(-1, 1)$ and standard variations $(1, 2)$. The results for MAVE, SADE with known score and for SADE with unknown score are shown on the Figure 11. We can see, that both SADE with known and unknown scores lose in case of low dimension of data, but more resistant to the curse of dimensionality (as expected as MAVE relies on non-parametric estimation in a space of dimension k, which is here larger). Moreover, the complexity of our moment-matching technique is linear in the number of observations n, while MAVE is superlinear.

6. Conclusion

In this paper we consider a general non-linear regression model and the dependence on a unknown k-dimensional subspace assumption. Our goal was direct

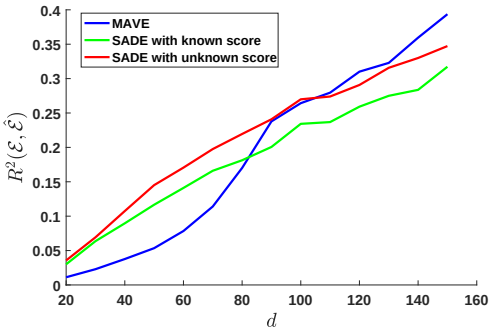

FIG 11. *Increasing dimensions, with rational model 5.3,* $n = 10000$*,* $k = 10$*,* $d = 20$ *to 150.*

estimation of this unknown k-dimensional space, which is often called the effective dimension reduction or e.d.r. space. We proposed new approaches (SADE and SPHD), combining two existing techniques (sliced inverse regression (SIR) and score function-based estimation). We obtained consistent estimation for $k > 1$ only using the first-order score and proposed explicit approaches to learn the score from data.

It would be interesting to extend our sliced extensions to learning neural networks: indeed, our work focused on the subspace spanned by w and cannot identify individual columns, while [20] showed that by using proper tensor decomposition algorithms and second-order score functions, columns of w can be consistently estimated (in polynomial time).

Acknowledgements

This work was funded by the MacSeNet Innovative Training Network. We would like to thank Anastasia Podosinnikova and Anatoli Juditsky for interesting discussions related to this work.

References

[1] A. Hyvärinen, *Estimation of non-normalized statistical models by score matching*, Journal of Machine Learning Research, 6 (2005), p. 695709.

[2] G. Arfken, *Divergence*, in Mathematical Methods for Physicists, Academic Press, Orlando, FL, 1985, ch. 1.7, pp. 37–42.

[3] A. Argyriou, T. Evgeniou, and M. Pontil, *Convex multi-task feature learning*, Machine Learning, 73 (2008), pp. 243–272.

[4] A. Argyriou, T. Evgeniou, and M. Pontil, *Convex multi-task feature learning*, Machine Learning, 73 (2008), pp. 243–272.

[5] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford University Press, 2013.

[6] D. R. Brillinger, *A Generalized Linear Model with 'Gaussian' Regressor Variables*, in A Festschrift for Erich L. Lehmann, K. D. P.J. Bickel and J. Hodges, eds., Woodsworth International Group, Belmont, California, 1982.

[7] S. Cambanis, S. Huang, and G. Simons, *On the Theory of Elliptically Contoured Distributions*, Journal of Multivariate Analysis, 11(3) (1981), pp. 368–385.

[8] R. D. Cook, *Save: a method for dimension reduction and graphics in regression*, Communications in Statistics - Theory and Methods, 29 (2000), pp. 2109–2121.

[9] R. D. Cook and H. Lee, *Dimension Reduction in Binary Response Regression*, Journal of the American Statistical Association, 94 (1999), pp. 1187–1200.

[10] R. D. Cook and S. Weisberg, *Discussion of 'Sliced Inverse Regression' by K. C. Li*, Journal of the American Statistical Association, 86 (1991), pp. 328–332.

[11] A. S. Dalalyan, A. Juditsky, and V. Spokoiny, *A new algorithm for estimating the effective dimension-reduction subspace*, Journal of Machine Learning Research, 9 (2008), pp. 1647–1678.

[12] N. Duan and K.-C. Li, *Slicing regression: a link-free regression method*, The Annals of Statistics, 19 (1991), pp. 505–530.

[13] K. Fukumizu, F. R. Bach, and M. I. Jordan, *Kernel dimension reduction in regression*, The Annals of Statistics, 37 (2009), pp. 1871–1905.

[14] L. Gyrfi, M. Kohler, A. Krzyzak, and H. Walk, *A distribution-free theory of nonparametric regression*, Springer series in statistics, Springer, New York, 2002.

[15] J. Hooper, *Simultaneous Equations and Canonical Correlation Theory*, Econometrica, 27 (1959), pp. 245–256.

[16] M. Hristache, A. Juditsky, and V. Spokoiny, *Direct estimation of the index coefficient in a single index model*, The Annals of Statistics, 29(3) (2001), pp. 595–623.

[17] T. Hsing and R. J. Carroll, *An asymptotic theory for sliced inverse regression*, The Annals of Statistics, 20(2) (1992), pp. 1040–1061.

[18] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, vol. 46, John Wiley & Sons, 2004.
[19] M. Janzamin, H. Sedghi, and A. Anandkumar, *Score function features for discriminative learning: Matrix and tensor framework*, CoRR, abs/1412.2863 (2014).
[20] ——, *Generalization Bounds for Neural Networks through Tensor Factorization*, CoRR, abs/1506.08473 (2015).
[21] K.-C. Li, *Sliced Inverse Regression for Dimensional Reduction*, Journal of the American Statistical Association, 86 (1991), pp. 316–327.
[22] ——, *On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma*, Journal of the American Statistical Association, 87 (1992), pp. 1025–1039.
[23] K.-C. Li and N. Duan, *Regression analysis under link violation*, The Annals of Statistics, 17 (1989), p. 10091052.
[24] Q. Lin, Z. Zhao, and J. S. Liu, *On consistency and sparsity for sliced inverse regression in high dimensions*, The Annals of Statistics, 46 (2018), pp. 580–610.
[25] M. McDonald, A. M .and Pontil and S. Stamos, *Spectral k-support norm regularization*, in Advances in Neural Information Processing Systems, 2014.
[26] C. Stein, *Estimation of the Mean of a Multivariate Normal Distribution*, The Annals of Statistics, 9 (1981), pp. 1135–1151.
[27] G. Stewart and J.-G. Sun, *Matrix perturbation theory (computer science and scientific computing)*, 1990.
[28] T. Stoker, *Consistent estimation of scaled coefficients*, Econometrica, 54 (1986), p. 14611481.
[29] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, Springer, 2009.
[30] V. Q. Vu, J. Lei, et al., *Minimax sparse principal subspace estimation in high dimensions*, The Annals of Statistics, 41 (2013), pp. 2905–2947.
[31] H. Wang and Y. Xia, *On directional regression for dimension reduction*, in J. Amer. Statist. Ass, Citeseer, 2007.
[32] ——, *Sliced regression for dimension reduction*, Journal of the American Statistical Association, 103 (2008), pp. 811–821.
[33] J. W.F. Donoghue, *Monotone Matrix Functions and Analytic Continuation*, Springer, 1974.
[34] Y. Xia, H. Tong, W. Li, and L.-X. Zhu, *An adaptive estimation of dimension reduction space*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64 (2002), pp. 363–410.
[35] Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu, *An adaptive estimation of dimension reduction space*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64 (2002), pp. 363–410.
[36] S. S. Yang, *General distribution theory of the concomitants of order statistics*, The Annals of Statistics, 5 (1977), pp. 996–1002.
[37] Y. Yu, T. Wang, R. J. Samworth, et al., *A useful variant of the davis–kahan theorem for statisticians*, Biometrika, 102 (2015), pp. 315–323.

[38] M. Yuan, *On the identifiability of additive index models*, Statistica Sinica, 21 (2011), pp. 1901–1911.
[39] L.-X. Zhu and K. W. Ng, *Asymptotics of sliced inverse regression*, Statistica Sinica, 5 (1995), pp. 727–736.

Appendix A: Appendix. Proofs

In this appendix, we provide proofs which were omitted in the main paper.

A.1. Probabilistic lemma

Let us first formulate and proof an auxiliary lemma about tail inequalities:

Lemma 5. *Let X be a non-negative random variable such that for some positive constants A and B, and all p:*

$$\mathbb{E}X^p \leqslant (A\sqrt{p} + Bn^{1/p}p^2)^p.$$

Then, for $t \geqslant (\log n)/2$:

$$\mathbb{P}(X \geqslant 3A\sqrt{t} + 3Bn^{1/t}t^2) \leqslant 3e^{-t}.$$

Proof. By Markov's inequality, for every non-negative integer p:

$$\mathbb{P}(X \geqslant 3A\sqrt{p} + 3Bn^{1/p}p^2) = \mathbb{P}(X^p \geqslant [3A\sqrt{p} + 3Bn^{1/p}p^2]^p) \leqslant$$

$$\leqslant \frac{[A\sqrt{p} + Bn^{1/p}p^2]^p}{[3A\sqrt{p} + 3Bn^{1/p}p^2]^p} \leqslant e^{-p}.$$

Consider any $t > (\log n)/2$ and $p = [t]$, then:

$$\mathbb{P}(X \geqslant 3A\sqrt{t} + 3Bn^{1/t}t^2) \leqslant \mathbb{P}(X \geqslant 3A\sqrt{p} + 3Bn^{1/p}p^2) \leqslant e^{-p} \leqslant 3e^{-t},$$

because function $f(t) = 3A\sqrt{t} + 3Bn^{1/t}t^2$ increases on $[(\log n)/2; +\infty)$. □

A.2. Proof of theorem 1

Proof. Let $\big(y_{(i)}, x_{(i)}\big)$, $i = 1, \ldots, n$ be the ordered data set, where $y_{(1)} \leqslant y_{(2)} \leqslant \ldots \leqslant y_{(n)}$. The $x_{(i)}$ are called the *concomitants* of order statistics by Yang [36]. Introduce double subscripts: $\ell_{(h,1)} = \ell\big(x_{(2h-1)}\big)$, $\ell_{(h,2)} = \ell\big(x_{(2h)}\big)$, $y_{(h,1)} = y_{(2h-1)}$ and $y_{(h,2)} = y_{(2h)}$.
Introduce firstly alternative matrix (under the weak assumptoin of its existance):

$$\mathcal{V}_{1,\mathbb{E}} = \mathbb{E}\big[\mathrm{Cov}(\mathcal{S}_1(x)|y)\big].$$

We have the following estimator for this matrix, for $c = 2$, and $H = n/c = n/2$:

$$\hat{\mathcal{V}}_{1,\mathbb{E}} = \frac{1}{n}\sum_{h=1}^{H}\big(\ell_{(h,1)} - \ell_{(h,2)}\big)\big(\ell_{(h,1)} - \ell_{(h,2)}\big)^\top,$$

Firstly, we estimate a norm of $\hat{\mathcal{V}}_{1,\mathbb{E}} - \mathcal{V}_{1,\mathbb{E}}$ and afterwards a norm of $\hat{\mathcal{V}}_{1,\mathrm{cov}} - \mathcal{V}_{1,\mathrm{cov}}$.

Thus, the deviation from the population version, may be split into four terms as follows:

$$\begin{aligned}
\hat{\mathcal{V}}_{1,\mathbb{E}} - \mathcal{V}_{1,\mathbb{E}} &= \frac{1}{n}\sum_{h=1}^{H}\big(\ell_{(h,1)} - \ell_{(h,2)}\big)\big(\ell_{(h,1)} - \ell_{(h,2)}\big)^\top - \mathbb{E}\eta\eta^\top \\
&= \frac{1}{n}\sum_{h=1}^{H}\big(\eta_{(h,1)} - \eta_{(h,2)}\big)\big(\eta_{(h,1)} - \eta_{(h,2)}\big)^\top - \mathbb{E}\eta\eta^\top \\
&\quad + \frac{1}{n}\sum_{h=1}^{H}\big(m(y_{(h,1)}) - m(y_{(h,2)})\big)\big(\eta_{(h,1)} - \eta_{(h,2)}\big)^\top \\
&\quad + \frac{1}{n}\sum_{h=1}^{H}\big(\eta_{(h,1)} - \eta_{(h,2)}\big)\big(m(y_{(h,1)}) - m(y_{(h,2)})\big)^\top \\
&\quad + \frac{1}{n}\sum_{h=1}^{H}\big(m(y_{(h,1)}) - m(y_{(h,2)})\big)\big(m(y_{(h,1)}) - m(y_{(h,2)})\big)^\top \\
&= T_4 + T_3 + T_2 + T_1.
\end{aligned}$$

We now bound each term separately.

Bounding T_1. We have

$$\begin{aligned}
T_1 &\preccurlyeq \frac{1}{n}\sum_{h=1}^{H} L^2 (y_{(h,1)} - y_{(h,2)})(y_{(h,1)} - y_{(h,2)})^\top \\
\mathrm{tr}T_1 = \|T_1\|_* &\leqslant \frac{1}{n}\sum_{h=1}^{H} L^2 \cdot \mathrm{diameter}(y_1,\dots,y_n)|y_{(h,1)} - y_{(h,2)}| \\
&\leqslant \frac{1}{n} L^2 \cdot \mathrm{diameter}(y_1,\dots,y_n)^2.
\end{aligned}$$

The range cannot grow too much, i.e., as $\log n$. Indeed, assuming without loss of generality that $\mathbb{E}y = 0$, we have $\max\{y_1,\dots,y_n\} \leqslant u/2$ and $\min\{y_1,\dots,y_n\} \geqslant -u/2$ implies that the range is less than u, and thus, $\mathbb{P}(\mathrm{diameter}(y_1,\dots,y_n) \geqslant u) \leqslant \mathbb{P}(\max\{y_1,\dots,y_n\} \geqslant u/2) + \mathbb{P}(\min\{y_1,\dots,y_n\} \leqslant -u/2) \leqslant n\mathbb{P}(y > u/2) + n\mathbb{P}(y < -u/2) \leqslant 2n\exp(-u^2/8\tau_y^2)$ by using sub-Gaussianity. Then, by selecting $u^2/8\tau_y^2 = \log(2n) + \log(8/\delta)$, with get with probability greater then $1 - \delta/8$ that $\mathrm{diameter}(y_1,\dots,y_n) \leqslant 2\sqrt{2}\tau_y\sqrt{\log(2n) + \log(8/\delta)}$.

Bounding T_2 and T_3. We also have

$$\begin{aligned}
\max\{\|T_2\|_*, \|T_3\|_*\} &\leqslant \frac{1}{n}\sum_{h=1}^{H} L|y_{(h,1)}y_{(h,2)}| \cdot \mathrm{diameter}(\eta_1,\dots,\eta_n) \\
&\leqslant \frac{1}{n} L \cdot \mathrm{diameter}(y_1,\dots,y_n) \cdot \mathrm{diameter}(\eta_1,\dots,\eta_n).
\end{aligned}$$

Like for T_1, the ranges cannot grow too much, i.e., as $\log n$. Similarly $\mathbb{P}(\text{diameter}((\eta_j)_1,\dots,(\eta_j)_n) \geqslant u) \leqslant n\mathbb{P}((\eta_j) > u/2) + n\mathbb{P}((\eta_j) < -u/2) \leqslant 2n\exp(-u^2/8\tau_\eta^2)$. We thus with get with probability greater then $1-\delta/(8d)$ that $\max_{j\in\{1,\dots,d\}} \text{diameter}((\eta_j)_1,\dots,(\eta_j)_n) \leqslant 2\sqrt{2}\tau_\eta\sqrt{\log(2n)+\log(8d/\delta)}$.

Thus combining the two terms above, with probability greater than $1-\delta/4$,

$$\|T_1\|_* + \|T_2\|_* + \|T_3\|_* \leqslant \frac{8L(L\tau_y^2 + 2\tau_\eta\tau_y\sqrt{d})}{n}\big(\log(2n)+\log(8d)+\log(1/\delta)\big).$$

Note the term in $\sqrt{d}$, which corresponds to the definition of the diameter $\text{diameter}(\eta_1,\dots,\eta_n)$ in terms of the ℓ_2-norm.

Bounding T_4. We have:

$$\begin{aligned} T_4 &= \frac{1}{n}\sum_{h=1}^{H}\big\{\eta_{(h,1)}\eta_{(h,1)}^\top + \eta_{(h,2)}\eta_{(h,2)}^\top - \eta_{(h,2)}\eta_{(h,1)}^\top - \eta_{(h,1)}\eta_{(h,2)}^\top\big\} - \mathbb{E}\eta\eta^\top \\ &= \frac{1}{n}\sum_{h=1}^{H}\big\{ - \eta_{(h,2)}\eta_{(h,1)}^\top - \eta_{(h,1)}\eta_{(h,2)}^\top\big\} + \frac{1}{n}\sum_{i=1}^{n}\eta_i\eta_i^\top - \mathbb{E}\eta\eta^\top = T_{4,1} + T_{4,2}. \end{aligned}$$

For the second term $T_{4,2}$ above, if we select any element indexed by a,b, then

$$\frac{1}{n}\sum_{i=1}^{n}(\eta_i)_a(\eta_i)_b - \mathbb{E}\eta_a\eta_b.$$

Using [5, Theorem 2.1], we get

$$\mathbb{E}[(\eta_i)_a(\eta_i)_b]^2 \leqslant \sqrt{\mathbb{E}(\eta_i)_a^4\mathbb{E}(\eta_i)_b^4} \leqslant 4(2\tau_\eta^2)^2 = 16\tau_\eta^4,$$

and

$$\mathbb{E}[|(\eta_i)_a(\eta_i)_b|^q] \leqslant \sqrt{\mathbb{E}(\eta_i)_a^{2q}\mathbb{E}(\eta_i)_b^{2q}} \leqslant 2q!(2\tau_\eta^2)^q = \frac{q!}{2}(2\tau_\eta^2)^{q-2}16\tau_\eta^4.$$

We can then use Bernstein's inequality [5, Theorem 2.10], to get that with probability less than e^{-t} then

$$\frac{1}{n}\sum_{i=1}^{n}(\eta_i)_a(\eta_i)_b - \mathbb{E}\eta_a\eta_b \geqslant 2\frac{\tau_\eta^2}{n}t + \sqrt{32\tau_\eta^4}\sqrt{t}/\sqrt{n}.$$

We can also get upper bound for this quantity, using $-\eta_a$ instead of η_a. Thus, with $t = \log\frac{8d^2}{\delta}$, we get that all $d(d+1)/2$ absolute deviations are less than $2\tau_\eta^2\left(\frac{\log\frac{8d^2}{\delta}}{n} + \sqrt{\frac{2\log\frac{8d^2}{\delta}}{n}}\right)$, with probability greater than $1-\delta/4$. This implies that the nuclear norm of the second term is less than $2d\sqrt{d}\tau_\eta^2\left(\frac{\log\frac{8d^2}{\delta}}{n} +\right.$

$\sqrt{\frac{2\log\frac{8d^2}{\delta}}{n}}\Big)$, because for any matrix $K \in \mathbb{R}^{d\times d}$: $\|K\|_* \leqslant d\sqrt{d}\|K\|_\infty$, where $\|K\|_\infty = \max_{j,k}|K_{jk}|$.

For the first term, we consider $Z = \sum_{h=1}^{H} (\eta_{(h,2)})_a(\eta_{(h,1)})_b$, and consider conditioning on $\mathbf{Y} = (y_1, \dots, y_n)$. A key result from the theory of order statistics is that the n random variables $\eta_{(h,2)}$, $\eta_{(h,1)}$, $h \in \{1, \dots, n/2\}$ are independent given Y [36]. This allows us to compute expectations.

Using Rosenthal's inequality [5, Theorem 15.11] conditioned on $\mathbf{Y}$, for which we have $\mathbb{E}((\eta_{(h,2)})_a(\eta_{(h,1)})_b|\mathbf{Y}) = 0$, we get:

$$\big[\mathbb{E}(|Z|^p|\mathbf{Y})\big]^{1/p} \leqslant$$

$$\leqslant \sqrt{8p}\big[\sum_h \mathbb{E}\big[((\eta_{(h,2)})_a^2(\eta_{(h,1)})_b^2|\mathbf{Y}\big]\big]^{1/2} + p\cdot 2\big[\mathbb{E}\max_h\big[((\eta_{(h,2)})_a^p(\eta_{(h,1)})_b^p|\mathbf{Y}\big]\big]^{1/p}$$
$$\leqslant \sqrt{8p}\big[\sum_h \mathbb{E}\big[((\eta_{(h,2)})_a^2(\eta_{(h,1)})_b^2|\mathbf{Y}\big]\big]^{1/2} + p\cdot 2\big[\sum_h \mathbb{E}\big[((\eta_{(h,2)})_a^p(\eta_{(h,1)})_b^p|\mathbf{Y}\big]\big]^{1/p}.$$

By taking the p-th power, we get:

$$\begin{aligned}\mathbb{E}(|Z|^p|\mathbf{Y}) &\leqslant 2^{p-1}\sqrt{8p}^p\big[\sum_h \mathbb{E}\big[((\eta_{(h,2)})_a^2(\eta_{(h,1)})_b^2|\mathbf{Y}\big]\big]^{p/2}\\ &+ 2^{p-1}p^p\cdot 2^p\sum_h \mathbb{E}\big[((\eta_{(h,2)})_a^p(\eta_{(h,1)})_b^p|\mathbf{Y}\big].\end{aligned}$$

By now taking expectations with respect to $\mathbf{Y}$, we get, using Jensen's inequality:

$$\begin{aligned}\mathbb{E}|Z|^p &\leqslant 2^{p-1}\sqrt{8p}^p\mathbb{E}\Big(\big[\sum_h(\eta_{(h,2)})_a^2(\eta_{(h,1)})_b^2\big]^{p/2}\Big)\\ &+ 2^{p-1}p^p\cdot 2^p\sum_h\mathbb{E}\big[((\eta_{(h,2)})_a^p(\eta_{(h,1)})_b^p\big]\\ &\leqslant 2^{p-1}\sqrt{8p}^p\mathbb{E}\Big(\big[\frac{1}{2}\sum_h(\eta_{(h,2)})_a^4 + (\eta_{(h,1)})_b^4\big]^{p/2}\Big)\\ &+ 2^{p-1}p^p\cdot 2^p\cdot\frac{1}{2}\sum_h\mathbb{E}\big[((\eta_{(h,2)})_a^{2p} + (\eta_{(h,1)})_b^{2p}\big]\\ &\leqslant 2^{p-1}\sqrt{8p}^p\mathbb{E}\Big(\big[\sum_i(\eta_i)_a^4 + (\eta_i)_b^4\big]^{p/2}\Big)\\ &+ 2^{p-1}p^p\cdot 2^p\sum_i\mathbb{E}\big[((\eta_i)_a^{2p} + (\eta_i)_b^{2p}\big].\end{aligned}$$

Because summing over all order statistics is equivalent to summing over all

elements. Thus, using the bound on moments of $(\eta_i)_b^2$, we get:

$$\begin{aligned}\mathbb{E}|Z|^p &\leqslant 2^{p-1}\sqrt{8p}^p 2^{p/2-1}\mathbb{E}\Big(\big[\sum_i (\eta_i)_a^4\big]^{p/2}\Big)\\ &+ 2^{p-1}\sqrt{8p}^p 2^{p/2-1}\mathbb{E}\Big(\big[\sum_i (\eta_i)_b^4\big]^{p/2}\Big) + 2^{p-1}p^p\cdot 2^p n\cdot 4p!(2\tau_\eta^2)^p\end{aligned}$$

We can now use [5, Theorem 15.10], to get

$$\begin{aligned}\Big[\mathbb{E}\Big(\big[\sum_i (\eta_i)_a^4\big]^{p/2}\Big)\Big]^{2/p} &\leqslant 2\mathbb{E}\big[\sum_i (\eta_i)_a^4\big] + \frac{p}{2}\big(\mathbb{E}\big[\max_i((\eta_i)_a^4)^{p/2}\big]\big)^{2/p}\\ &\leqslant 2\mathbb{E}\big[\sum_i (\eta_i)_a^4\big] + \frac{p}{2}\big(\mathbb{E}\big[\sum_i((\eta_i)_a^4)^{p/2}\big]\big)^{2/p}\\ &\leqslant 2n\times 4(2\tau_\eta^2)^2 + \frac{p}{2}n^{2/p}\mathbb{E}\eta_i^{2p}\\ &\leqslant \big(32n + n^{2/p}\frac{p}{2}(2p!)^{2/p}\big)\tau_\eta^4\\ &\leqslant \big(32n + n^{2/p}p^3\big)\tau_\eta^4.\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}|Z|^p &\leqslant 2^p\sqrt{8p}^p 2^{p/2-1}\big(32n + n^{2/p}p^3\big)^{p/2}\tau_\eta^{2p} + 2^{p-1}p^p\cdot 2^p n\cdot 4p!(2\tau_\eta^2)^p\\ &\leqslant 2^{3p-1}p^{p/2}\cdot 2^{p/2-1}\big((32n)^{p/2} + n\cdot p^{3p/2}\big)\tau_\eta^{2p} + 2^{3p+1}\tau_\eta^{2p}np^{2p}\\ &\leqslant \big(2^{6p-2}p^{p/2}n^{p/2} + np^{2p}[2^{7p/2-2} + 2^{3p+1}]\big)\tau_\eta^{2p}\\ &\leqslant \Big(64^p\cdot p^{p/2}n^{p/2} + 19^p\cdot np^{2p}\Big)\tau_\eta^{2p}.\end{aligned}$$

Thus

$$(\mathbb{E}|Z|^p)^{1/p} \leqslant \Big(64\cdot\sqrt{p}n^{1/2} + 19\cdot n^{1/p}p^2\Big)\tau_\eta^2.$$

Thus, for any $\delta\leqslant 1/n$, using Lemma 5 for random variable Z/n with $t = \log(\frac{12d^2}{\delta}) > (\log n)/2$ and we obtain:

$$\mathbb{P}\left[\left|\frac{Z}{n}\right| \geqslant \frac{192\tau_\eta^2}{\sqrt{n}}\sqrt{t} + \frac{57\tau_\eta^2}{n}n^{1/t}t^2\right] \leqslant 3e^{-t} \Rightarrow$$

$$\mathbb{P}\left[\left|\frac{Z}{n}\right| \geqslant \frac{192\tau_\eta^2}{\sqrt{n}}\sqrt{\log(\tfrac{12d^2}{\delta})} + \frac{57\tau_\eta^2}{n}n^{1/\log(\frac{12d^2}{\delta})}\log^2(\tfrac{12d^2}{\delta})\right] \leqslant \frac{\delta}{4d^2} \Rightarrow$$

$$\mathbb{P}\left[\left|\frac{Z}{n}\right| \geqslant \frac{192\tau_\eta^2}{\sqrt{n}}\sqrt{\log(\tfrac{12d^2}{\delta})} + \frac{155\tau_\eta^2}{n}\log^2(\tfrac{12d^2}{\delta})\right] \leqslant \frac{\delta}{4d^2}.$$

Combining all terms $T_1, T_2, T_3, T_{4,1}$ and $T_{4,2}$ we get with probability not less than $1-\delta$:

$$\|\hat{\mathcal{V}}_{1,\mathbb{E}}-\mathcal{V}_{1,\mathbb{E}}\|_* \leqslant \frac{1}{n}\cdot\left[8L(L\tau_y^2+2\tau_\eta\tau_y\sqrt{d})\cdot\log(\tfrac{16dn}{\delta})+2d\sqrt{d}\tau_\eta^2\log\tfrac{8d^2}{\delta}+155\tau_\eta^2 d\log^2(\tfrac{12d^2}{\delta})\right]$$
$$+\frac{1}{\sqrt{n}}\left[2d\sqrt{d}\tau_\eta^2\sqrt{2\log\tfrac{8d^2}{\delta}}+192\tau_\eta^2 d\sqrt{\log\tfrac{12d^2}{\delta}}\right].$$

Rearranging terms and replacing δ by $\delta/2$, with probability not less, than $1-\delta/2$:

$$\|\hat{\mathcal{V}}_{1,\mathbb{E}}-\mathcal{V}_{1,\mathbb{E}}\|_* \leqslant \frac{195d\sqrt{d}\tau_\eta^2}{\sqrt{n}}\sqrt{\log\frac{24d^2}{\delta}}+\frac{8L^2\tau_y^2+16\tau_\eta\tau_y L\sqrt{d}+157\tau_\eta^2 d\sqrt{d}}{n}\log^2\frac{32d^2n}{\delta}.$$

Using expression:

$$\mathcal{V}_{1,\mathrm{cov}}+\mathcal{V}_{1,\mathbb{E}}=\mathrm{cov}[\ell_1(x)],$$

we can suggest estimator for $\mathcal{V}_{1,\mathrm{cov}}$ as $\hat{\mathcal{V}}_{1,\mathrm{cov}}=\frac{1}{n}\sum\limits_{i=1}^n \ell(x_i)\ell(x_i)^\top-\hat{\mathcal{V}}_{1,\mathbb{E}}$.

Applying the triangle inequality:

$$\|\hat{\mathcal{V}}_{1,\mathrm{cov}}-\mathcal{V}_{1,\mathrm{cov}}\|_* \leqslant \|\hat{\mathcal{V}}_{1,\mathbb{E}}-\mathcal{V}_{1,\mathbb{E}}\|_*+\|\frac{1}{n}\sum_{i=1}^n \ell(x_i)\ell(x_i)^\top-\mathbb{E}(\ell(x)\ell(x)^\top)\|_*.$$

To estimate the second term, we can use the same arguments as for bounding $T_{4,2}$: with probability greater than $1-\delta/2$, $\|\frac{1}{n}\sum\limits_{i=1}^n \ell(x_i)\ell(x_i)^\top-\mathbb{E}(\ell(x)\ell(x)^\top)\|_*$ is less than $2d\sqrt{d}\tau_\ell^2\left(\frac{\log\frac{4d^2}{\delta}}{n}+\sqrt{\frac{2\log\frac{4d^2}{\delta}}{n}}\right)$. Finally, combining this bound with bound for $\|\hat{\mathcal{V}}_{1,\mathrm{cov}}-\mathcal{V}_{1,\mathrm{cov}}\|_*$:

$$\|\hat{\mathcal{V}}_{1,\mathrm{cov}}-\mathcal{V}_{1,\mathrm{cov}}\|_* \leqslant \frac{d\sqrt{d}(195\tau_\eta^2+2\tau_\ell^2)}{\sqrt{n}}\sqrt{\log\frac{24d^2}{\delta}}+$$
$$\frac{8L^2\tau_y^2+16\tau_\eta\tau_y L\sqrt{d}+(157\tau_\eta^2+2\tau_\ell^2)d\sqrt{d}}{n}\log^2\frac{32d^2n}{\delta}.$$

□

A.3. Proof of Theorem 2

Proof. Using triangle inequality, we get:

$$\|\mathcal{V}_{1,\mathrm{cov}}(\theta^*)-\hat{\mathcal{V}}_{1,\mathrm{cov}}(\hat{\theta})\|_* \leqslant$$
$$\|\mathcal{V}_{1,\mathrm{cov}}(\theta^*)-\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta^*)\|_*+\|\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta^*)-\hat{\mathcal{V}}_{1,\mathrm{cov}}(\hat{\theta})\|_*=\mathcal{F}_1+\mathcal{F}_2. \quad \text{(A.1)}$$

Theorem 1 supplies us with non-asymptotic analysis for the $\mathcal{F}_1$ term: with probability not less then $1-\delta/2$:

$$\|\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta^*) - \mathcal{V}_{1,\mathrm{cov}}(\theta^*)\|_* \leqslant \frac{d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2)}{\sqrt{n}}\sqrt{\log\frac{48d^2}{\delta}} + \frac{8L^2\tau_y^2 + 16\tau_\eta\tau_y L\sqrt{d} + (157\tau_\eta^2 + 2\tau_\ell^2)d\sqrt{d}}{n}\log^2\frac{64d^2n}{\delta}. \tag{A.2}$$

For the second term, let us firstly analyse the norm $\|\theta^* - \hat{\theta}\|$. For simplicity, introduce notation: $\hat{C} = \frac{1}{n}\sum_{i=1}^n \Psi(x_i)\Psi(x_i)^\top \in \mathbb{R}^{m\times m}$, $C = \mathbb{E}\big[\Psi(x)\Psi(x)^\top\big] \in \mathbb{R}^{m\times m}$, $\hat{b} = \frac{1}{n}\sum_{i=1}^n(\nabla\cdot\Psi)(x_i) \in \mathbb{R}^m$ and $b = \mathbb{E}\big[(\nabla\cdot\Psi)(x)\big] \in \mathbb{R}^m$

Let us estimate $\|\hat{C} - C\|_F$ and $\|\hat{b} - b\|$.
Introduce notation:

$$C^c_{a,b} = \frac{1}{n}\sum_{i=1}^n \Psi^a_c(x_i)\cdot\Psi^b_c(x_i) - \mathbb{E}\big[\Psi^a_c(x)\cdot\Psi^b_c(x)\big].$$

It can be shown like in the bounding of $T_{4,2}$ term in the Theorem's 1 proof, using [5, Theorem 2.1] and Bernstein's inequality [5, Theorem 2.10] that with probability less then e^{-t}:

$$C^c_{a,b} \geqslant 2\frac{\tau_\Psi^2}{n}t + \sqrt{32\tau_\Psi^4}\sqrt{t}/\sqrt{n}.$$

Taking $t = \log\frac{2m^2d}{\delta}$ we show that:

$$\mathbb{P}\left[\|\hat{C} - C\|_F \geqslant 2\tau_\Psi^2\sqrt{m^3d}\left(\frac{\log\frac{2m^2d}{\delta}}{n} + \sqrt{\frac{2\log\frac{2m^2d}{\delta}}{n}}\right)\right] < \delta.$$

According to assumption **(M2)** and Hoeffding bound:

$$\mathbb{P}\left[\|\hat{b}_i - b_i\| \geqslant \frac{t}{n}\right] \leqslant e^{\frac{-t^2}{2\tau^2_{\nabla\Psi}n}},$$

combining inequalities for all m components of b, we have:

$$\mathbb{P}\left[\|b - \hat{b}\| \geqslant \frac{\tau_{\nabla\Psi}\sqrt{m}}{\sqrt{n}}\sqrt{\log\tfrac{m^2}{\delta^2}}\right] \leqslant \delta.$$

Now, let estimate $\|\theta^* - \hat{\theta}\|$:

$$\theta^* - \hat{\theta} = \hat{C}^{-1}\hat{b} - C^{-1}b = \hat{C}^{-1}(\hat{b} - b) + (\hat{C}^{-1} - C^{-1})b \Rightarrow$$

$$\|\theta^* - \hat{\theta}\| \leqslant \frac{\|\hat{b} - b\|}{\lambda_{\min}(\hat{C})} + \frac{\|b\|_2\cdot\|C - \hat{C}\|_{\mathrm{op}}}{\lambda_{\min}(C)\cdot\lambda_{\min}(\hat{C})}.$$

For n large enough (for some constants c_1 and c_2, not depending on n and δ: $n > c_1 + c_2 \log \frac{1}{\delta}$) $\|\hat{C} - C\| \leqslant \dfrac{\lambda_{\min}}{2}$ with probability more, than $1 - \delta/12$ hence for this n: $\lambda_{\min}(\hat{C}) \geqslant \dfrac{\lambda_{\min}(C)}{2}$, hence, combining 3 estimations for $\|\hat{b} - b\|$, $\|\hat{C} - C\|$ and for $\lambda_{\min}(\hat{C})$, we obtain estimation for $\|\theta^* - \hat{\theta}\|$: with probability not less, than $1 - \delta/4$:

$$\|\theta^* - \hat{\theta}\| \leqslant$$

$$\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\sqrt{n}\lambda_{\min}(C)}\sqrt{2\log\tfrac{12m}{\delta}} + \frac{2\|b\|}{\lambda^2_{\min}(C)} \cdot 2\tau^2_{\Psi}\sqrt{m^3 d}\left(\frac{\log\frac{24m^2 d}{\delta}}{n} + \sqrt{\frac{2\log\frac{24m^2 d}{\delta}}{n}}\right) \tag{A.3}$$

Consider now (a, b) element of $\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\mathrm{cov}}(\hat{\theta})$ (using 4.2):

$$\left[\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\mathrm{cov}}(\hat{\theta})\right]_{a,b} = \frac{1}{n}\sum_{i,j=1}^{n}\frac{\alpha_{i,j}}{|I_h(i,j)| - 1}\sum_{\alpha,\beta=1}^{m}\Psi(x_i)^{\alpha}_{a}\Psi(x_j)^{\beta}_{b}\cdot|\theta^*_{\alpha}\theta^*_{\beta} - \hat{\theta}_{\alpha}\hat{\theta}_{\beta}| \leqslant$$

$$\frac{1}{2n}\sum_{i,j=1}^{n}\frac{\alpha_{i,j}}{|I_h(i,j)| - 1}\sum_{\alpha,\beta=1}^{m}\left[\left[\Psi(x_i)^{\alpha}_{a}\right]^2 + \left[\Psi(x_j)^{\beta}_{b}\right]^2\right]\cdot|\theta^*_{\alpha}\theta^*_{\beta} - \hat{\theta}_{\alpha}\hat{\theta}_{\beta}| \leqslant$$

$$\sum_{\alpha,\beta=1}^{m}\frac{1}{2n}\left[\sum_{i=1}^{n}\left[\Psi(x_i)^{\alpha}_{a}\right]^2 + \sum_{i=1}^{n}\left[\Psi(x_i)^{\beta}_{b}\right]^2\right]\cdot\left|\theta^*_{\alpha}\theta^*_{\beta} - \hat{\theta}_{\alpha}\hat{\theta}_{\beta}\right|,$$

because every row of binary matrix $\left\{\alpha_{i,j}\right\}_{i,j=\overline{1,n}}$ has exactly $|I_h(i,j)| - 1$ non-zero elements.

Now, let us estimate desired norm:

$$\|\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\mathrm{cov}}(\hat{\theta})\|_* \leqslant m\cdot\sum_{k=1}^{n}\sum_{l=1}^{d}\frac{\sum\limits_{i=1}^{n}|\Psi^k_l(x_i)|^2}{n}\cdot\left(2\|\theta^*\| + \|\hat{\theta} - \theta^*\|\right)\cdot\|\hat{\theta} - \theta^*\| \tag{A.4}$$

Using again [5, Theorem 2.1] and Bernstein's inequality [5, Theorem 2.10] with probability not less then $1 - \delta/8$:

$$\sum_{k=1}^{n}\sum_{l=1}^{d}\frac{\sum\limits_{i=1}^{n}|\Psi^k_l(x_i)|^2}{n} \leqslant \mathbb{E}\|\Psi(x)\|^2_2 + 2md\tau^2_{\psi}\left(\frac{\log\frac{16md}{\delta}}{n} + \sqrt{\frac{2\log\frac{16md}{\delta}}{n}}\right) \tag{A.5}$$

For n large enough (for some constants c_1 and c_2, not depending on n and δ: $n > c_1 + c_2 \log \frac{1}{\delta}$):

$$\sum_{k=1}^{n}\sum_{l=1}^{d}\frac{\sum\limits_{i=1}^{n}|\Psi^k_l(x_i)|^2}{n} \leqslant \frac{3}{2}\mathbb{E}\|\Psi(x)\|^2_2 \text{ with probability no less then } 1 - \delta/8,$$

and

$$\big(2\|\theta^*\| + \|\hat{\theta} - \theta^*\|\big) \leqslant 3\|\theta^*\| \quad \text{with probability no less then } 1 - \delta/8$$

$$\|\hat{\mathcal{V}}_{1,\mathrm{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\mathrm{cov}}(\hat{\theta})\|_* \leqslant m \cdot \frac{9}{2}\mathbb{E}\|\Psi(x)\|_2^2 \cdot \|\theta^*\| \cdot$$

$$\left[\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\sqrt{n}\lambda_{\min}(C)}\sqrt{2\log\tfrac{12m}{\delta}} + \frac{2\|b\|}{\lambda_{\min}^2(C)} \cdot 2\tau_\Psi^2\sqrt{m^3 d}\left(\frac{\log\frac{24m^2d}{\delta}}{n} + \sqrt{\frac{2\log\frac{24m^2d}{\delta}}{n}}\right)\right]$$

with probability not less then $1 - \delta/2$.

Combining and simplifying obtained inequality and (A.2) we obtain final inequality: with probablility not less, then $1 - \delta$:

$$\|\mathcal{V}_{1,\mathrm{cov}}(\theta^*) - \hat{\mathcal{V}}_{1,\mathrm{cov}}(\hat{\theta})\|_* \leqslant \frac{1}{\sqrt{n}}\sqrt{2\log\frac{48m^2d}{\delta}} \times$$

$$\left[d\sqrt{d}(195\tau_\eta^2 + 2\tau_\ell^2) + \frac{9m}{2}\mathbb{E}\|\Psi(x)\|_2^2 \cdot \|\theta^*\|\left(\frac{2\tau_{\nabla\Psi}\sqrt{m}}{\lambda_0} + \frac{4\|b\|\tau_\Psi^2\sqrt{m^3 d}}{\lambda_0}\right)\right] + O\left(\frac{1}{n}\right).$$

□